

How Much Work Can AI Agents Do?

Task Value Inside the AI Time-Horizon Frontier

Tom Harris*

October 2026

Abstract

How much work can AI agents perform autonomously? A key predictor of whether an AI agent can complete a task is how long the task takes a human, and on the benchmarks of METR (Model Evaluation and Threat Research) the time horizon of frontier models has doubled roughly every four months since 2023. I combine METR’s agent success curves with human task-duration estimates from the Anthropic Economic Index across 18,796 tasks in O*NET, the US Occupational Information Network, weighting each task’s share of the US wage bill by the probability of autonomous completion that METR’s curves imply. Tasks accounting for 46 percent of the US wage bill are expected to succeed under the latest frontier model, more than three times the 14 percent at the start of 2025. Coverage of individual jobs is broad but partial: no occupation reaches 90 percent. Unlike exposure rubrics, which ask whether AI could speed up a task, or usage data, which conflate capability with adoption, the measure gives macroeconomic models a capability-based value for the share of work AI could automate, one that moves with the frontier.

JEL codes: J24, O33, J23, E24. **Keywords:** AI agents, time horizons, task-based models, automation, labor markets.

1 Introduction

METR (Model Evaluation and Threat Research) measures the performance of AI agents by the length of the tasks they can complete autonomously, expressed as the time a human would need for the same work (Kwa et al., 2025). On METR’s evaluation suite, this time horizon has doubled roughly every 4.2 months since 2023 (METR, 2026), and its chart is now followed closely for signs that progress is speeding up or slowing down (New York Times, 2026). Over the same period, the Anthropic Economic Index (AEI) has tracked which work tasks users delegate to Claude and estimated how long that work would take a person without AI (Handa et al., 2025; Anthropic, 2026). This paper places both measures on a single scale, US work tasks weighted by their share of the wage bill, to measure two quantities: the share of the wage bill in tasks a frontier agent is expected to complete, and the share in tasks that workers actually delegate to an AI system.

*Department of Economics, London School of Economics and Political Science; t.m.harris@lse.ac.uk. I thank Xavier Jaravel, Danial Lashkari and Yechan Park for helpful comments. This draft is preliminary; comments and corrections are very welcome.

Macroeconomic models of AI rely on assumed values for both of these quantities. In growth models, the effects of AI turn on the share of tasks it automates (Trammell and Korinek, 2026). In the scenario framework of Korinek et al. (2026), the “extreme” 2030 scenario, for example, assumes that AI can affect 50 percent of all tasks (the affected mass m), that it is used on 60 percent of the instances of those tasks (the diffusion share d), and that nine in ten AI-performed tasks are automated rather than augmented (the automation share ψ). Across the values chosen for these parameters, outcomes range from negligible disruption to 17.9 percent cognitive unemployment, 15.4 percent annual GDP growth and a labor share of 45.2 percent. The anchor used so far for the affected mass comes from usage data, which mixes what AI can do with how far its use has spread. This paper provides an empirical value that rests on capability instead: it uses a *capability curve* to measure the share of work AI could perform outright, and traces how that share has grown. Any task an agent can complete alone is one that AI affects, so the measure is also a lower bound on the affected mass. If METR’s curves carry over to the rest of the economy, this lower bound already exceeds the affected mass assumed in the substantial 2030 scenario, so the spread of outcomes depends mainly on diffusion, the gain per task and the automation share.

To measure capability, I use METR’s fitted logistic curves, which map an agent’s probability of completing a task to the logarithm of the task’s human time. In the language of task-based models (Acemoglu and Restrepo, 2018), the curve orders tasks by the time they take a person and makes the frontier’s time horizon an automation threshold that moves out at a measured rate. Rather than counting only the tasks below a threshold, I sum the wage bill weighted by each task’s probability of success given its length. This expected-success share is a continuous version of the threshold measure. Existing measures of occupational exposure to AI instead judge from written descriptions of work whether AI is relevant to it, without observing AI attempt it, and rank occupations at one point in time (Webb, 2020; Felten et al., 2021; Eloundou et al., 2024). The METR curve is measured on agents attempting tasks, gives a probability of success rather than a rank, and has a trend that can be projected forward. The two disagree: among tasks within an agent’s scope, occupations that rank high on the exposure score of Eloundou et al. (2024) tend to rank low on expected success (rank correlation -0.41).

I anchor the estimates on two models: Claude Opus 4.6, the most capable model inside METR’s 16-hour reliability range, and Mythos Preview (early), a pre-release checkpoint that is the latest measured frontier but sits above that range. Tasks carrying 40 percent of the US wage bill are expected to succeed under Claude Opus 4.6 and 46 percent under Mythos Preview (early), up from 14 percent under o1, the frontier at the start of 2025. The level depends heavily on which tasks count as within an agent’s scope; the growth does not. Under every scope rule I consider, the share has grown by a factor of 3 or more since then, and faster still under stricter standards of success. Nearly all work in scope is now shorter than the 50 percent horizon, so further gains depend more on how reliably agents complete short tasks than on how long a task they can attempt. Across jobs, coverage is partial: 19 percent of US employment is in occupations where at least half of the work is expected to succeed, and no occupation reaches 90 percent. Tasks delegated to Claude carry 25 to 30 percent of the wage bill in each AEI data window and 49 percent across all windows, and 18 of the 46 points of expected success are in tasks with no recorded use. Where AI is used, tasks an agent is more likely to complete are more often handed over outright than worked on together.

These shares use METR’s definition of success, in which an agent completes a well-specified task that is scored automatically. Stricter standards give lower success: when repository maintainers judge whether they would merge an agent’s code, METR finds horizons about 7 times shorter (Whitfill et al., 2026), and at that standard the expected-success share is 27 percent, equivalent to

the frontier of about 12 months earlier.

Applying METR’s curves outside software is the central assumption of this paper. Success may follow different curves in different domains (Kwa and Cheng, 2025), and the broadest evidence on how it does so outside software comes from Mertens et al. (2026). Workers experienced in each occupation graded responses from 41 language models to 6,648 tasks in O*NET, the US Occupational Information Network, drawn from 22 occupation groups, from management and finance to healthcare, education and law. Each response was judged on whether a manager would accept it without edits. By the third quarter of 2025, frontier models met that standard about 70 percent of the time on tasks of one to four hours, and success falls only slowly with task length. For a given model, their curve predicts lower success than METR’s on tasks shorter than about an hour and higher success on longer ones. At the release of GPT-5, the share of the wage bill in tasks that their curve expects to succeed is 41 percent, against 30 percent on METR’s curves. The best evidence from outside software therefore does not suggest that METR’s curves overstate the level of what AI can do (Section 5).

The closest work is the January 2026 Anthropic Economic Index report (Appel et al., 2026), which combines success rates classified from Claude conversations with task time shares into an “effective AI coverage” of occupations; Massenkoff and McCrory (2026), who build from the same usage data the observed-exposure measure that Korinek et al. (2026) use to anchor their scenarios; and Mertens et al. (2026), who grade model responses to O*NET tasks and project success forward to 2030. This paper differs in using a capability curve measured on agents rather than success classified from usage or graded single responses, in weighting by the wage bill, and in tracing the measure across frontier models from 2023 to the present. It builds on task-based models of technical change (Autor et al., 2003; Acemoglu and Restrepo, 2018) and on measures of occupational exposure to AI (Brynjolfsson et al., 2018; Webb, 2020; Felten et al., 2021; Eloundou et al., 2024), which Acemoglu (2025) turns into a macroeconomic estimate by combining exposure with the share of exposed tasks that is cost-effective to automate (Svanberg et al., 2024). This paper also provides macroeconomic models of AI with a value for the affected mass that rests on *capability* rather than *use*, and a record of how fast that value has grown.

2 Data and Measurement

2.1 Success curves from METR

METR’s evaluation suite covers 228 computer-based tasks in 79 families. Most come from HCAST, a set of software engineering, machine learning, cybersecurity and research tasks with completion times benchmarked on human contractors (Rein et al., 2025). For each model, METR fits a logistic curve of success on the $\log_2$ of a task’s human time, and the 50 percent time horizon is the length at which the curve crosses one half (Kwa et al., 2025). I rebuilt every run from METR’s per-task results (Time Horizon 1.1, May 2026) and refit the curves with METR’s method. The refit reproduces METR’s published horizons to within 0.02 percent for every model and supplies the coefficients of the continuous curve (Appendix B). METR flags horizons above 16 hours as unreliable because its suite has few long tasks. Mythos Preview (early) is the only frontier model above that threshold, and Claude Opus 4.6 is the last model measured below it; I report both. The frontier at a date is the model with the longest 50 percent horizon released by then.

2.2 Task durations from the AEI

The AEI reports the share of Claude conversations mapped to each O*NET task, alongside a model-generated estimate of the time a person would need to complete the delegated work without AI (Handa et al., 2025; Anthropic, 2026). I take durations from the June 2026 release, which maps conversations to exact O*NET task IDs. Usage comes from seven Claude.ai data windows from December 2024 to May 2026 and from the windows for Anthropic’s application programming interface (API) published since September 2025. Durations are observed for 2,814 tasks, carrying 26 percent of the wage bill. For the remaining tasks I impute the geometric mean of observed durations among tasks that share an O*NET detailed work activity, backing off to broader activities where no peer is observed (Appendix A); leave-one-out checks show no bias at the level of detailed work activities (Appendix Table A9). The wage-weighted median duration is 3.1 hours.

2.3 Wage weights

I give each O*NET task a share of its occupation’s wage bill, from Occupational Employment and Wage Statistics (OEWS) May 2025 employment and mean annual wages (U.S. Bureau of Labor Statistics, 2026; National Center for O*NET Development, 2026), in proportion to its importance among the occupation’s core tasks. Summed over all 18,796 tasks, this gives \$10.72 trillion across 153.4 million jobs. Weighting by importance and relevance, or admitting O*NET’s supplemental tasks, moves the expected-success share by less than a point; hours-implied weights lower it by about two points (Appendix Table A8). Every estimate comes from one file with a row per O*NET task; an interactive version, with every task drawn against the frontier, is at <https://aitasks-ecru.vercel.app/>.

2.4 Tasks in scope

A task in this paper is an O*NET task statement, such as “record patients’ medical information and vital signs” for registered nurses, and it carries the wage weight. Its length is the AEI estimate of the human time for the piece of it that users delegate in a conversation, while success is measured on METR’s tasks, which are self-contained and well specified. These units need not coincide (Section 5).

Which work a software agent could do is a matter of definition. The baseline excludes a task if *any* of its detailed work activities maps to a physical generalized work activity, such as handling objects, operating vehicles or caring for others, or to one of the nine generalized work activities that O*NET defines as live interaction with people: establishing relationships, selling or influencing, resolving conflicts, working directly with the public, building teams, teaching, directing subordinates, coaching and staffing. Activities that O*NET defines without reference to a mode, such as communicating with colleagues “by telephone, in written form, e-mail, or in person”, stay in scope. The strict rule reflects the premise that the measure counts the completion of a task as defined: a software agent cannot complete a mixed task unless an employer unbundles it, and METR’s tasks are self-contained deliverables, which a meeting or the supervision of staff is not. Excluded tasks carry 40 percent of the wage bill, a little over half of it in physical tasks and the rest in interactive ones, and are assigned a success probability of zero.

The rule records what a task involves, not where it is done. For the results on occupations in Section 3.4, I therefore also exclude every task of an occupation that cannot be done remotely under the rule of Dingel and Neiman (2020), rebuilt on O*NET 30.2, or that requires close physical

contact, such as school teaching (Appendix K). Panel B of Table 1 reports the expected-success share under alternative rules, and Appendix C gives examples.

2.5 The expected-success share

Let V_t denote the wage bill attached to task t , D_t its duration in human hours, and $P_d(D)$ the fitted success curve of the frontier model at date d . The expected-success share is

$$S_d = \frac{\sum_t V_t P_d(D_t) \mathbf{1}[t \text{ in scope}]}{\sum_t V_t}.$$

The threshold shares, F_d^{50} and F_d^{80} , replace $P_d(D_t)$ with an indicator equal to one if D_t is below the 50 or 80 percent horizon. S_d instead integrates the full curve, giving partial credit to tasks beyond the threshold and discounting shorter tasks by their residual failure rate. It is a measure of capability: it does not account for the cost of checking the agent’s work or for the friction of reassigning standing responsibilities to an agent.

Evaluating S_d across the economy requires extrapolating METR’s success curves out of domain. The core assumption is that, at a given human duration, an agent’s probability of completing a paralegal’s task in scope equals the probability predicted for a software engineering task of the same length. Duration is a natural index of difficulty because a longer task chains more steps, each of which can fail (Ord, 2025); the assumption is that the rate of failure per step is the same outside software. Within METR’s suite, task length predicts average success well. With four to six runs per task, binomial noise caps the attainable R^2 of task success rates at roughly 0.57, and length accounts for about 43 percent of the explainable variation for Mythos Preview (early) (Appendix D). The measure is linear in P_d , so it requires only that the curve capture mean success at a given length, not the outcome of each task. Dropping any one task family moves the share by no more than two points.

Outside the suite, the evidence comes from less controlled measurements, and it points to flatter curves. Success rates classified from Claude conversations fall more slowly with task length than METR’s curves, but they reflect which tasks users choose to delegate and how they iterate on them (Appel et al., 2026). Expert grading of one-shot deliverables also gives a flatter gradient (Patwardhan et al., 2025), and METR finds horizons up to two orders of magnitude shorter on tasks that require operating a computer through its screen, although they grow at similar rates (Kwa and Cheng, 2025). The broadest evidence, workers’ evaluations of model responses to 6,648 O*NET tasks across occupations (Mertens et al., 2026), gives a curve that is both flatter and higher than METR’s: it predicts lower success on tasks of less than an hour but higher success on longer ones, where most of the wage bill lies. On this paper’s tasks it gives a larger expected-success share than METR’s curves at every date in their sample, so on this evidence METR’s curves are, if anything, conservative about the level of the share, though not about its growth (Section 5). I keep METR’s curves, whose gradient is measured on autonomous completion under controlled conditions, and report an in-domain check on Computer and Mathematical occupations (Standard Occupational Classification group 15, SOC 15), where the assumption is least strained (Appendix E).

3 Results

3.1 Where the wage bill sits

Figure 1 sets METR’s curves above the US wage bill by task duration. Most of the wage bill falls between 15 minutes and 16 hours, and nearly all of it in scope lies below 12.0 hours, the shorter of the two 50 percent horizons, so the 50 percent threshold share is close to its ceiling under either model. This near-saturation reflects the AEI duration data, not a ceiling on capability: AEI durations measure modular pieces of work delegated in chat, which are short by selection. Being shorter than the 50 percent horizon means only that the curve gives better-than-even odds: tasks in scope have a wage-weighted expected success of 68 percent under Claude Opus 4.6 and 78 percent under Mythos Preview (early), not 100 percent.

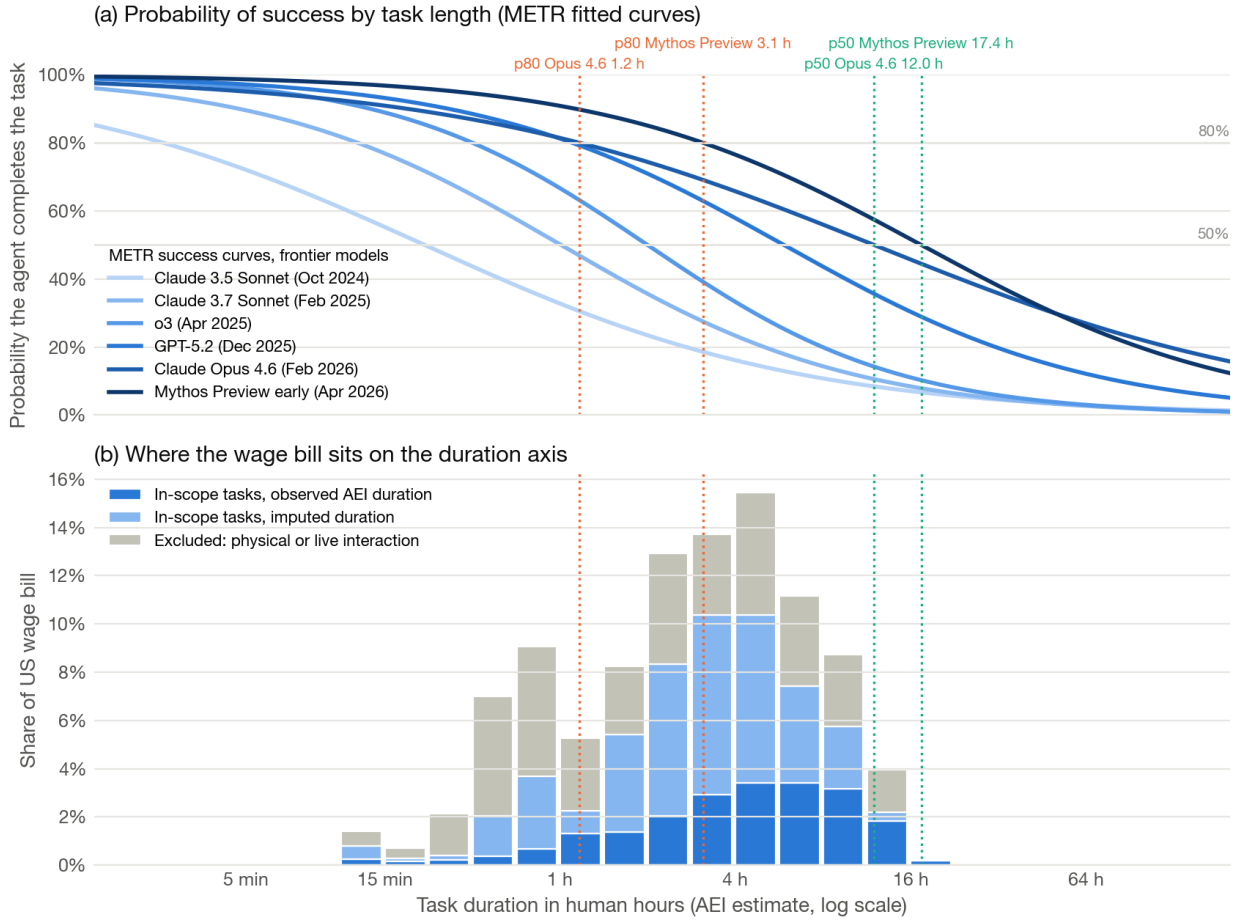


Figure 1: Success curves and the wage bill by task length. (a) METR’s fitted success curves for six frontier models. (b) The US wage bill by task duration on the same axis, in half-log₂ bins. Dotted lines: the 50 and 80 percent horizons of Claude Opus 4.6 and Mythos Preview (early). The two 80 percent horizons bracket 19 percent of the wage bill.

Weighting by those probabilities, tasks carrying 40 percent of the US wage bill are expected to succeed under Claude Opus 4.6 and 46 percent under Mythos Preview (early) (Table 1, panel A). A compound bootstrap that resamples METR’s curve, adds measurement noise to AEI durations and reruns the imputation gives a 95 percent interval of 40 to 51 percent under the latest frontier; it

reflects sampling uncertainty only. The share in tasks shorter than the 80 percent horizon rises from 8 to 27 percent between the two models, because that horizon moves from 1.2 to 3.1 hours across the densest part of the duration distribution. The 80 percent horizon of Mythos Preview (early) is poorly measured, which makes the integrated share the more stable aggregate (Appendix G).

3.2 Growth

Figure 2(a) traces the expected-success share as the frontier advanced. It grew from 14 percent under o1 at the start of 2025 to 46 percent under Mythos Preview (early). Table 1 shows that the level of the share depends on definitions while its growth does not. Under alternative rules for which tasks are in scope, the share under Mythos Preview (early) ranges from 24 percent, counting only occupations that can be done remotely without close contact, to 60 percent, keeping live interaction in scope: the definition of automatable work moves the share by more than a model generation. Under every rule, however, the share has grown by a factor of 3 or more since the start of 2025. Under stricter standards of success it has grown faster still, by a factor of 7 at the maintainers' standard, because a shorter horizon puts more of the wage bill on the steep part of the curve.

The share depends on the curve only through its 50 percent horizon and its slope, so it can be read off for any future METR estimate (Appendix Figure A7). At the slope of Mythos Preview (early), one more doubling of the horizon raises the share from 46 to 51 percent. Nearly all work in scope is already shorter than the 50 percent horizon, so most of the remaining gap to the ceiling of 60 percent, the in-scope share of the wage bill, is the curve's failure rate on those tasks. At the median slope, the share reaches 50 percent at a horizon of about 28 hours, roughly 3 months beyond Mythos Preview (early) at the current doubling rate, and then approaches its ceiling slowly. Further gains therefore depend more on how reliably agents complete relatively short tasks than on the length of task they can attempt.

Table 1: The expected-success share of the US wage bill (percent) under alternative definitions. Columns: the frontier model at the first AEI data window (o1), the last model inside METR’s reliable range (Claude Opus 4.6) and the latest measured frontier (Mythos Preview (early)). Rows are shares of the whole US wage bill, except the second, whose denominator is the wage bill of tasks in scope, and the SOC 15 row, whose denominator is that group’s wage bill. Growth: Mythos Preview (early) over o1. Parentheses: 95 percent intervals from 500 draws of a compound bootstrap of METR’s curve, AEI durations and the imputation; ^a METR’s curve only. Panel B changes which tasks count as in scope (Section 2.4). Panel C divides the 50 percent horizon by the factor shown, holding the slope: maintainers’ judgment of whether they would merge an agent’s code (Whitfill et al., 2026), and the ratio of contractors’ to maintainers’ completion times (Kwa et al., 2025). The last row subtracts checking time equal to a quarter of the task’s human time.

	o1 Dec 2024	Opus 4.6 Feb 2026	Mythos Preview Apr 2026	Growth (times)
<i>A. Baseline</i>				
Expected-success share	14 (10–18) ^a	40 (34–46)	46 (40–51)	3.3
as a share of tasks in scope	24	68	78	3.3
Share in tasks shorter than the 80 percent horizon	0	8	27	—
Computer and mathematical occupations (SOC 15)	14	54	63	4.3
<i>B. Which tasks are in scope</i>				
Physical filter only (live interaction in scope)	17	52	60	3.4
Lenient (excluded only if all activities are)	16	45	52	3.3
Selling kept in scope	14	41	47	3.3
Remote occupations without close contact only	6	21	24	3.7
<i>C. Standard of success and unit of time</i>				
Maintainers’ standard (horizon ÷ 7)	4	25	27	6.9
All durations doubled	9	35	40	4.4
Experienced workers’ time (horizon ÷ 5 to 18)	2–5	17–27	18–31	6.1–9.1
Net of checking time of a quarter of the task	3	25	31	12.1

3.3 Capability and use

Figure 2(a) also sets capability beside use. Tasks with a published usage share on Claude.ai or the API carry 25 to 30 percent of the wage bill in each AEI data window, and 49 percent if every task that has appeared in any window is counted. That union is a generous measure of use: requiring a task to reach one in ten thousand conversations cuts it to 24 percent, and leaving out API traffic lowers it to 44 percent (Appendix Table A10). Use is not a subset of capability. At the 80 percent horizon of Mythos Preview (early), 23 percent of the wage bill is beyond the frontier but appears in Claude usage, so workers use AI on tasks an agent cannot yet complete reliably alone, while 12 percent is inside the frontier but unused. Measured with the full curve, 29 of the 46 points of expected success are in tasks that have appeared in Claude usage and 18 in tasks that have not; of the 49 percent of the wage bill in use, 37 points are in scope. A further 12 percent of the wage bill is in excluded tasks that nonetheless appear in Claude usage, most of it in tasks excluded for live interaction: users bring Claude the preparation and follow-up, not the meeting itself (Appendix Figure A5).

The gap between capability and use is therefore not a simple adoption lag. The usage measure includes augmentation beyond the frontier and excludes unused capability within it. Short tasks may also go unused because handing work to an agent has a fixed cost, in specifying the task and supplying its context, that is large relative to the time a short task saves (Appel et al., 2026). AEI

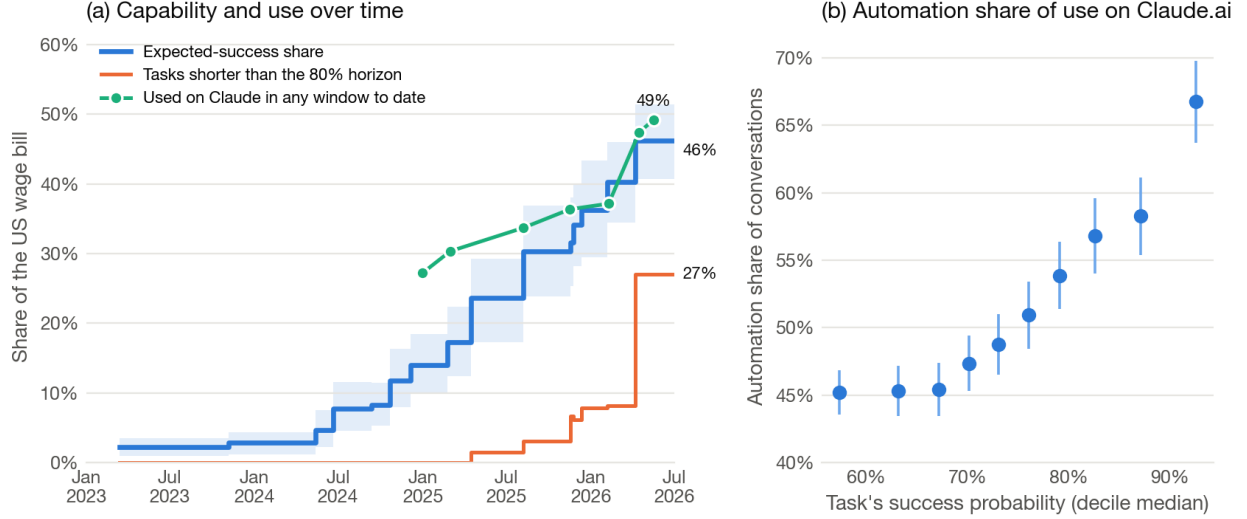


Figure 2: Capability and use. (a) Blue: expected-success share of the US wage bill under the frontier model at each date, the model with the longest 50 percent horizon released by then; band: 95 percent interval from a bootstrap of METR’s curve. Orange: share of the wage bill in tasks in scope shorter than the frontier’s 80 percent horizon. Green: share in tasks with a published Claude.ai or API usage share in any AEI data window to date. API data begin with the September 2025 window. (b) Automation share of Claude.ai conversations in May 2026, for the 2,378 tasks in scope with a published share, by decile of the task’s success probability under Mythos Preview (early); bars are 95 percent intervals. Automation: directive and feedback-loop conversations, as a share of conversations with a collaboration pattern (Handa et al., 2025).

data are also extensive-margin only: they record whether a task appears in conversations, not the intensity or the labor hours of the AI’s contribution.

Task length is also associated with how AI is used. The AEI counts as *automation* the conversations in which the user hands Claude a task to complete or only relays feedback such as error messages, and as *augmentation* those of iteration, learning and validation (Handa et al., 2025). Among the tasks in scope used on Claude.ai in May 2026, the automation share rises from 45 percent in the tenth of tasks with the lowest success probability to 67 percent in the highest (Figure 2(b)). Comparing tasks within the same occupation, a success probability one standard deviation higher goes with an automation share 6.3 points higher, against 2.8 points for the exposure score of Eloundou et al. (2024) in the same regression. Shorter tasks are thus more often handed over outright, as METR’s curve predicts. The same pattern would arise if short tasks were simply easier to hand over in a single exchange, whatever the agent’s ability; separating the two requires variation in capability at a given task length, such as the same task observed as the frontier moves.

3.4 From tasks to jobs

Figure 3(a) shows the expected-success share by occupation group. Computer, engineering, office, business and science occupations have the highest shares, at about 60 percent of their wage bill. Cleaning, food service, construction and healthcare support rank lowest, because of the physical filter. Management, sales, education and the arts lose most to the exclusion of live interaction, because much of their work is directing, teaching or selling to people.

The expected-success share also differs from rubric-based exposure measures. Across 782 occupations,

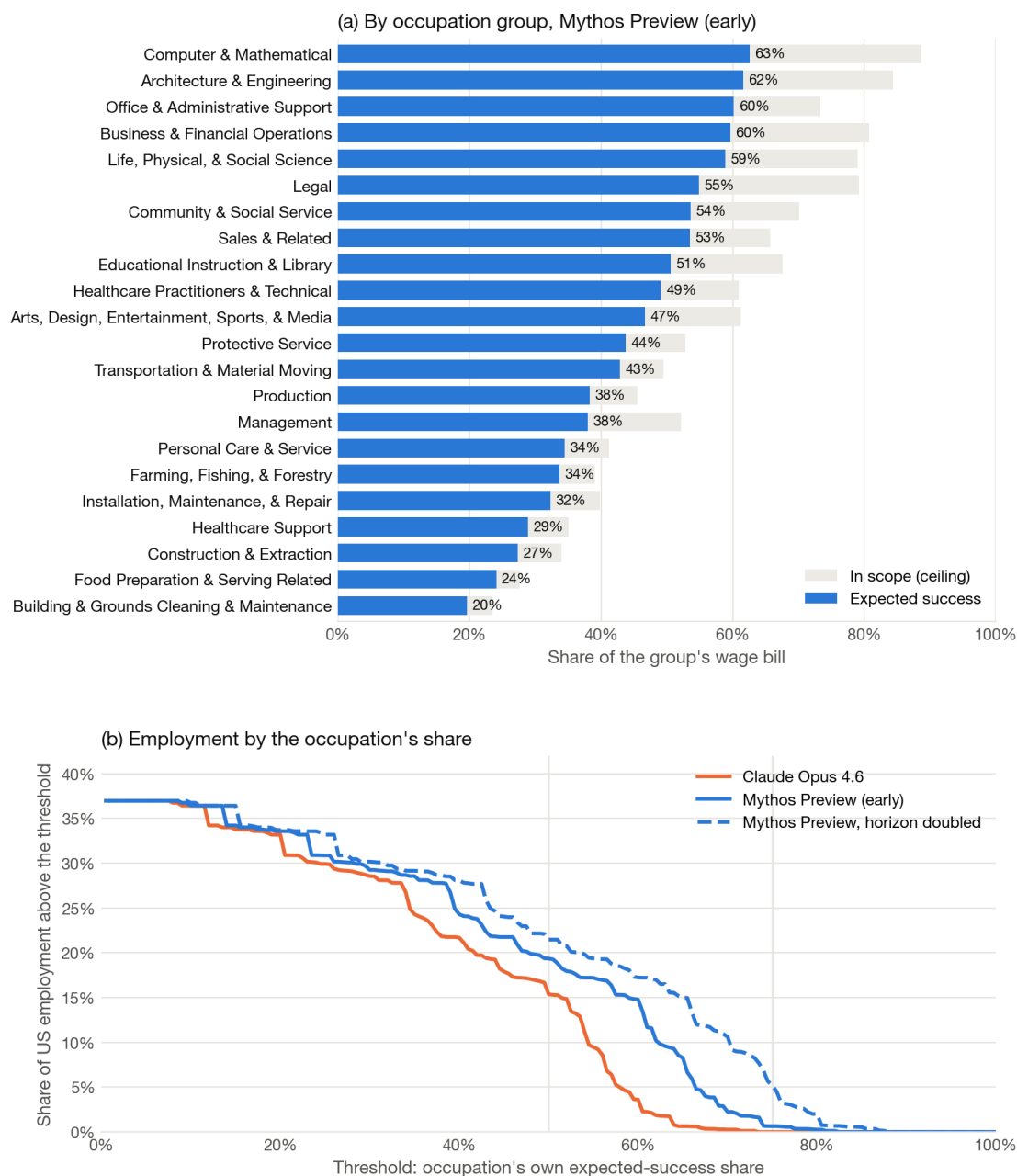


Figure 3: From tasks to jobs. (a) Expected-success share of each SOC major group's wage bill under Mythos Preview (early) (blue), against the group's in-scope share, its ceiling (grey). (b) Share of US employment in 6-digit occupations whose own expected-success share is at least the threshold. Excluded tasks, and every task of an occupation that cannot be done remotely (Dingel and Neiman, 2020) or needs close physical contact, count as zero, which is why the curves start below 40 percent. Dashed: the 50 percent horizon of Mythos Preview (early) doubled, at the same slope. Denominator: 153.4 million jobs.

its rank correlation with the value-weighted exposure score of [Eloundou et al. \(2024\)](#) is 0.75, but most of this agreement comes from the excluded set, which both measures largely respect; computing both over tasks in scope only, the rank correlation is -0.41 (Appendix I). The two measures answer different questions. The rubric asks whether a large language model could cut the time a person needs for a task by at least half, and does not consider how long the task is; the expected-success share asks whether an agent could complete the task alone, which becomes less likely as the task gets longer. Long analytical work, such as that of software developers (median task duration 8.1 hours), scores high on the first and lower on the second, while short clerical and front-line service work scores the other way. The long durations of software work partly reflect what users delegate: durations of coding-heavy tasks rose most when the AEI data began to include Cowork, Anthropic’s agentic desktop product.

An occupation is a bundle of tasks, so coverage of tasks does not translate directly into displacement from jobs. For each 6-digit occupation I compute its own expected-success share, counting as zero every task of an occupation that requires an in-person setting (Section 2.4). Under Mythos Preview (early), 19 percent of US employment is in occupations where at least half of the task value is expected to succeed, and 1 percent where at least three quarters is (Figure 3(b)). No occupation reaches 90 percent; the highest share is 85 percent. The ceiling comes from the success curve rather than from excluded duties: the top occupation has no excluded tasks, and under Mythos Preview (early) even a one-hour task succeeds with probability 0.91. The ceiling therefore moves with the frontier. One more doubling of the horizon would put 5 percent of employment in occupations where at least three quarters of task value is expected to succeed. The occupations with the highest shares are records, transcription and clerical occupations, with some mapping and science occupations, whose tasks are short; their durations are mostly imputed from related work activities, so the order at the top is not precise (Appendix Table A12).

Adding up task probabilities assumes that tasks can be unbundled and automated one at a time. If the uncovered tasks are instead the core of the job, high task-level coverage may change occupational demand very little until the entire bundle is covered ([Autor, 2015](#)). Linear aggregation also overstates displacement when tasks are complements ([Gans and Goldfarb, 2026](#)), and whether automation raises or lowers the value of the remaining work depends on whether the automated tasks are the expert or the inexpert ones ([Autor and Thompson, 2025](#)). The shares are also measured at the grain at which work is delegated: an agent that completes the short pieces of a long task leaves the work of combining them to a person, and O*NET does not record that work. Capability is also not the same as comparative advantage. Delegating has a fixed cost of specifying the task and checking the result, so the short clerical tasks that top the occupational ranking may be the least worth handing over. With a weakest-link aggregator, in which an occupation is only as exposed as the lowest success probability among its more important tasks, 3 percent of employment is above the 50 percent threshold. An expected-success share of 46 percent therefore means partial capability across many occupations rather than the complete displacement of a few. Whether partial coverage turns into substitution or augmentation depends on the cost of delegating, on whether a task has the scale to repay the fixed cost of adapting AI to it ([Lashkari et al., 2026](#)), and on how firms reorganize work, which are beyond the scope of this paper.

4 Implications for Macroeconomic Models

The model of Korinek et al. (2026) lets AI affect only cognitive work, which they define as twelve SOC major groups that account for 0.624 of employment, so the expected-success share I set against their affected mass m counts only tasks in those groups, divided by the whole wage bill as in their definition. It measures the work AI could perform outright, which their model calls automation, whether or not AI is yet used on it. In their notation it is closest to ψm when every instance is performed with AI ($d = 1$), and it is a lower bound on m measured as capability. Their mid-2026 anchor of 0.14 is built instead from Claude usage (Massenkoff and McCrory, 2026). It counts only tasks with substantial observed use, and it is low mainly for that reason: on this paper’s weights and in their cognitive groups, the same penetration measure gives 0.14, the share of the wage bill in tasks that appeared in Claude usage in May 2026 is 0.26, and the share in tasks that appeared in any window is 0.42. The expected-success share under Mythos Preview (early) is 0.36. It is not larger than every measure of use; what distinguishes it is that it measures what AI can do and moves with the frontier. If METR’s curves carry over to the rest of the economy, this lower bound already lies between the 0.3 and 0.5 that their substantial and extreme scenarios assume for m in 2030 (Appendix Table A15).

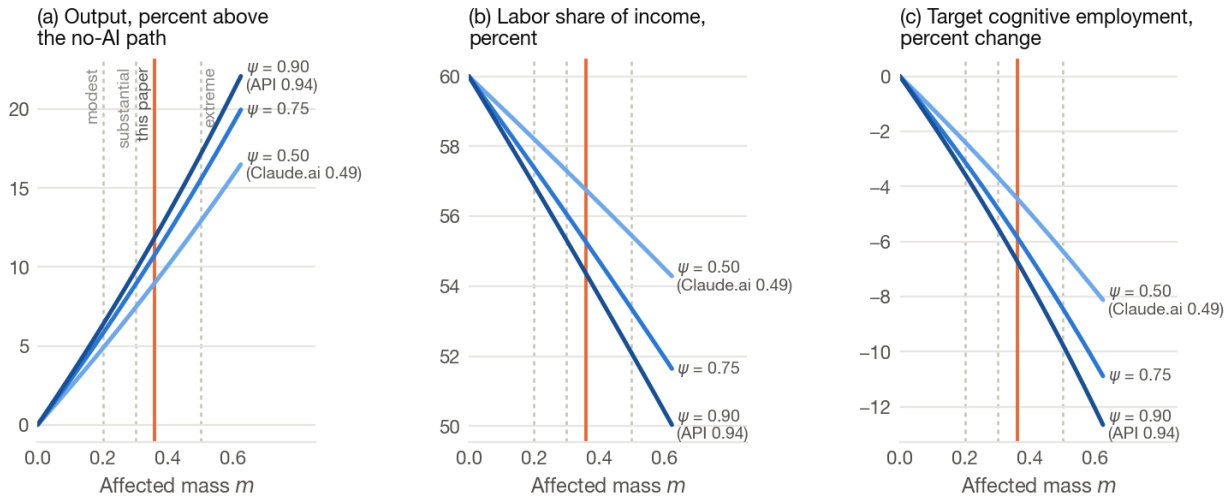


Figure 4: Static targets of the model of Korinek et al. (2026) as the affected mass m varies, for their three automation shares ψ . Diffusion (0.4), the gain per instance (0.45) and reinstatement (0.25) are at their substantial 2030 values; capital supply elasticity 3. Grey dotted lines: their 2030 values of m . Orange: this paper’s expected-success share under Mythos Preview (early) in the twelve occupation groups they classify as cognitive, as a share of the whole wage bill (0.36), a lower bound on m ; counting only occupations that can be done remotely and without close contact it is 0.23. Labels in brackets: the AEI’s measured automation shares, June 2026. (c) The level of employment cognitive occupations adjust toward, not employment in 2030.

Figure 4 solves the static part of their model, their Proposition 1, which gives the labor share, output and the employment each occupation group adjusts toward; at their parameter values it reproduces their published labor share for 2030 closely (Appendix M). The affected mass and the diffusion share enter the model only as a product, so the affected mass moves wages and employment only in proportion to diffusion. With capability already above their substantial value, the spread of their 2030 outcomes depends mainly on diffusion, the gain per task and the automation share. In their modest scenario the labor share falls by 0.6 points by 2030; with the affected mass set to the measured value and nothing else changed, it falls by 1.1 points, against 15 points in the

extreme scenario. Across tasks, the automation share is higher where the success probability is higher (Section 3.3), which fits their choice to pair a larger affected mass with a higher automation share. Its level depends more on how AI is deployed than on what it can do: about half of Claude.ai conversations are automation, against nine tenths of API traffic.

5 Limitations

Four limitations matter most. The first is the duration mismatch. METR measures human time for complete, well-specified tasks performed by contractors, whereas the AEI estimates time for the modular pieces of work users delegate in chat. Delegated pieces are short by selection, so AEI durations may understate the length of employer-defined tasks and bias the shares upward. Doubling every duration lowers the expected-success share under Mythos Preview (early) to 40 percent. METR’s human times also come from low-context contractors, who take 5 to 18 times longer than repository maintainers on the same issues (Kwa et al., 2025); if AEI durations instead reflect experienced workers in context, the share falls to 18–31 percent (Table 1, panel C). Claude’s time estimates correlate with actual task times on software tickets about as well as developers’ own estimates (Tamkin and McCrory, 2025), but only 26 percent of the wage bill has an observed duration. Hold-out checks show the imputation is unbiased at the level of detailed work activities, but not whether it carries over to tasks that are never delegated.

Second, the success curve is extrapolated beyond software. Within METR’s suite, the curve overstates success in the one-to-four-hour band, which inflates the headline share by about 2 points, within sampling noise (Appendix Table A3). Outside the suite, success curves derived from AEI usage, from expert grading on the GDPval benchmark and from worker evaluations across O*NET tasks are all flatter (Appendix E). On the flatter curve of Mertens et al. (2026), the share starts higher and grows far less: from 01 at the start of 2025 to GPT-5, the end of their sample, their release-date trend gives a share that rises from 37 to 41 percent of the wage bill, while METR’s curves give 14 and 30 percent. Their slope also varies across occupations, from 0.39 per tenfold increase in length in personal care to none in legal work, which a single curve ignores. Their curve measures single responses that workers judge usable without edits, on self-contained tasks with all context supplied, rather than work an agent carries out step by step. They note that supplying context in practice is costly, which is one reading of the gap between capability and use in Section 3.3.

Third, the level depends on what counts as success. At the maintainers’ standard the expected-success share is 27 percent. The gap is wider still for freelance work: on the Remote Labor Index, METR’s curves predict a 30 to 51 percent completion rate, yet clients accepted only 1.2 to 4.2 percent of the AI deliverables (Mazeika et al., 2025). I do not convert that gap into a share, because the benchmark reports one rate per model and 69 percent of its projects are in video, audio, design, CAD and game development (Appendix J). These gaps are far larger than the sampling interval. The exclusion of live interaction may also go too far: voice agents already answer telephones, and some teaching and supervision can be done in writing. Keeping selling in scope gives 47 percent, and the physical filter alone 60 percent.

Fourth, the measure gives a task the agent completes with probability 0.7 credit for 70 percent of its value, with no allowance for the time a worker spends checking the output. If checking an agent’s work takes a quarter of the task’s human time, the net share falls to 31 percent (Appendix L). Only 8.0 percent of the wage bill sits in tasks the curve expects to succeed nine times in ten, and

some tasks need higher reliability still before they are worth delegating (Kwa, 2026). Experienced developers in METR’s own trial were slower with AI than without it (Becker et al., 2025).

6 Conclusion

If METR’s curves carry over, the capability frontier covers in expectation just under half of the US wage bill, up from 14 percent at the start of 2025. Its level depends on which tasks count as within the scope of an agent, which the data cannot yet settle; its growth, by a factor of 3 or more since the start of 2025, holds under every definition I consider. Further gains depend more on how reliably agents complete relatively short tasks than on how long a task they can attempt. For now the coverage is broad but partial: the frontier reaches parts of most jobs that can be done remotely and the whole of few. For macroeconomic models of AI, the measure gives a capability-based value for the share of work AI could automate. At that value, the spread of outcomes in the scenarios of Korinek et al. (2026) depends mainly on how far AI use spreads, what it gains per task and how much of it is automation, and the first two have yet to be measured. Refining the estimates requires human completion times for tasks defined at the level of the occupation, and success curves measured outside software under a stated standard of success, such as those of Mertens et al. (2026).

References

- Daron Acemoglu. The simple macroeconomics of AI. *Economic Policy*, 40(121):13–58, 2025. doi: 10.1093/epolic/eiae042.
- Daron Acemoglu and Pascual Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6): 1488–1542, 2018.
- Anthropic. The Anthropic Economic Index: data releases. <https://huggingface.co/datasets/Anthropic/EconomicIndex>, 2026. Releases of February 2025 to June 2026.
- Ruth Appel, Maxim Massenkoff, Peter McCrory, Miles McCain, Ryan Heller, Tyler Neylon, and Alex Tamkin. The Anthropic Economic Index report: Economic primitives. Anthropic, 15 January 2026, January 2026. URL <https://www.anthropic.com/research/anthropic-economic-index-january-2026-report>.
- David Autor and Neil Thompson. Expertise. *Journal of the European Economic Association*, 23(4): 1203–1271, 2025. doi: 10.1093/jeea/jvaf023.
- David H. Autor. Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–30, 2015.
- David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333, 2003.
- Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein. Measuring the impact of early-2025 AI on experienced open-source developer productivity. arXiv preprint arXiv:2507.09089, 2025. URL <https://arxiv.org/abs/2507.09089>.

- Erik Brynjolfsson, Tom Mitchell, and Daniel Rock. What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, 108:43–47, 2018.
- Fabrizio Dell’Acqua, Edward McFowland III, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Kraye, François Cadelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2):403–423, 2026. doi: 10.1287/orsc.2025.21838. Previously circulated as Harvard Business School Working Paper 24-013 (2023).
- Jonathan I. Dingel and Brent Neiman. How many jobs can be done at home? *Journal of Public Economics*, 189:104235, 2020. doi: 10.1016/j.jpubeco.2020.104235.
- Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, 2024. doi: 10.1126/science.adj0998.
- Edward Felten, Manav Raj, and Robert Seamans. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12):2195–2217, 2021.
- Joshua S. Gans and Avi Goldfarb. O-Ring automation. Working Paper 34639, National Bureau of Economic Research, 2026. URL <https://www.nber.org/papers/w34639>.
- Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. Which economic tasks are performed with AI? Evidence from millions of Claude conversations. *arXiv preprint arXiv:2503.04761*, 2025.
- Anton Korinek, Charles I. Jones, Szymon Sacher, Tess Cotter, and Peter McCrory. Economic scenarios for transformative AI. Working Paper 2026-02, The Anthropic Institute, September 2026. URL <https://www-cdn.anthropic.com/files/4zrzovbb/website/cf58f84d46a4a76bf5a5b039ac695fba6b80041c.pdf>.
- Thomas Kwa. Clarifying limitations of time horizon. METR Notes, 22 January 2026, January 2026. URL <https://metr.org/notes/2026-01-22-time-horizon-limitations/>.
- Thomas Kwa and Vincent Cheng. How does time horizon vary across domains? METR Blog, 14 July 2025, July 2025. URL <https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/>.
- Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Chris Painter, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long software tasks. *arXiv preprint arXiv:2503.14499*, 2025.
- Danial Lashkari, Wensu Li, Christina Qiu, and Neil Thompson. AI, scale, and task-based theory of automation. Working paper, July 2026, 2026. URL <https://sites.google.com/view/danial-lashkari/1lqt>.

- Maxim Massenkoff and Peter McCrory. Labor market impacts of AI: A new measure and early evidence. Anthropic research report, 5 March 2026, March 2026. URL <https://www.anthropic.com/research/labor-market-impacts>.
- Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Schwag, Shivam Singhal, et al. Remote labor index: Measuring AI automation of remote work. arXiv preprint arXiv:2510.26787, 2025. URL <https://arxiv.org/abs/2510.26787>.
- Matthias Mertens, Adam Kuzee, Brittany S. Harris, Harry Lyu, Wensu Li, Jonathan Rosenfeld, Meiri Anto, Martin Fleming, and Neil Thompson. Crashing waves vs. rising tides: Findings on AI automation from thousands of worker evaluations of labor market tasks. arXiv preprint arXiv:2604.01363, July 2026. URL <https://arxiv.org/abs/2604.01363v3>. Version 3, 23 July 2026. Version 1 (1 April 2026) was titled “Preliminary Findings” and used about 17,000 evaluations.
- METR. Task-Completion Time Horizons of Frontier AI Models. <https://metr.org/time-horizons>, 2026. Time Horizon 1.1; data files `benchmark_results_1.1.yaml` and `task_results_1.1.yaml`, updated 8 May 2026.
- National Center for O*NET Development. O*NET 30.2 database. <https://www.onetcenter.org/database.html>, 2026.
- Toby Ord. Is there a half-life for the success rates of AI agents? arXiv preprint arXiv:2505.05115, 2025. URL <https://arxiv.org/abs/2505.05115>.
- Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, et al. GDPval: Evaluating AI model performance on real-world economically valuable tasks. arXiv preprint arXiv:2510.04374, 2025. URL <https://arxiv.org/abs/2510.04374>.
- David Rein, Joel Becker, Amy Deng, Seraphina Nix, Chris Canal, Daniel O’Connel, Pip Arnott, Ryan Bloom, Thomas Broadley, Katharyn Garcia, Brian Goodrich, Max Hasin, Sami Jawhar, Megan Kinniment, Thomas Kwa, Aron Lajko, Nate Rush, Lucas Jun Koba Sato, Sydney Von Arx, Ben West, Lawrence Chan, and Elizabeth Barnes. HCAST: Human-calibrated autonomy software tasks. *arXiv preprint arXiv:2503.17354*, 2025.
- Kevin Roose. How do you measure an A.I. boom? The New York Times, 17 April 2026, April 2026. URL <https://www.nytimes.com/2026/04/17/technology/how-do-you-measure-an-ai-boom.html>.
- Maja S. Svanberg, Wensu Li, Martin Fleming, Brian C. Goehring, and Neil C. Thompson. Beyond AI exposure: Which tasks are cost-effective to automate with computer vision? MIT FutureTech Working Paper, SSRN 4700751, 2024. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4700751.
- Alex Tamkin and Peter McCrory. Estimating AI productivity gains from Claude conversations. Anthropic research report, November 2025. URL <https://www.anthropic.com/research/estimating-productivity-gains>. Web page dated 25 November 2025; PDF dated 5 November 2025.
- Philip Trammell and Anton Korinek. Economic growth under transformative AI. *Annual Review of Economics*, 18(1):589–611, 2026. doi: 10.1146/annurev-economics-051624-061002.
- U.S. Bureau of Labor Statistics. Occupational employment and wage statistics, May 2025 national estimates. <https://www.bls.gov/oes/>, 2026.

Michael Webb. The impact of artificial intelligence on the labor market. Working paper, Stanford University, 2020.

Parker Whitfill, Cheryl Wu, Joel Becker, and Nate Rush. Many SWE-bench-passing PRs would not be merged into main. METR Notes, 10 March 2026, March 2026. URL <https://metr.org/notes/2026-03-10-many-swe-bench-passing-prs-would-not-be-merged-into-main/>.

Appendix

This appendix gives the data construction in full, the evidence on METR’s curves within and outside its suite, the exhibits and robustness checks behind the text, and the comparison with Korinek et al. (2026).

A Data Construction

The estimates in this paper are constructed from a single task-level dataset. Every aggregate reported in the text derives from an analysis file with one row per O*NET task, which records the task’s wage weight, observed or imputed AEI duration, physical classification, Eloundou et al. (2024) exposure label, usage flags and fitted success probabilities under the frontier models.

- **METR time horizons:** Data are downloaded from metr.org/time-horizons (benchmark_results_1_1.yaml and task_results_1_1.yaml, updated 8 May 2026). The curve refits follow METR’s public methodology (github.com/METR/eval-analysis-public): a weighted logistic regression of run success on $\log_2$ human minutes, with weights equal to the family weight divided by runs per task, normalized to one per model, using 10^{-5} regularization. Standard errors and intervals are generated via 500 bootstrap draws over family, task, and run. The frontier at date d is defined as the model with the longest 50 percent horizon released on or before d , and an AEI data window is paired with the frontier as of its last day. For the robust variants (Section F), the curves are refit using a probit link, logistic functions with asymptotes fixed at 0.85, 0.90, or 0.95, HCAST tasks only, and a leave-one-family-out procedure.
- **Anthropic Economic Index (AEI):** Sourced from huggingface.co/datasets/Anthropic/EconomicIndex. Usage shares are extracted from seven Claude.ai data windows spanning December 2024 to May 2026, and from the five API windows published from September 2025. Per-task human time (human_only_time_mean, which averages the estimated hours for a human to complete the delegated work without AI) uses the June 2026 release as the baseline, as it uniquely provides exact O*NET task IDs. Releases prior to June 2026 identify tasks by text string; I recover 87 to 91 percent of usage via exact string matching and drop the residual (mostly retired O*NET occupations). Missing durations are imputed using the simple mean over a task’s detailed work activities of the geometric mean of observed durations in each activity (36 percent of the wage bill). If no activity has an observed peer, the imputation backs off sequentially to the intermediate (22 percent) and generalized work activity (15 percent), and finally to the global geometric mean (0.8 percent). Tasks map to 1.27 detailed work activities on average.
- **O*NET 30.2:** Task statements, ratings (importance, relevance, frequency), and the work-activity hierarchy are merged to construct the wage weights. These weights use O*NET’s core tasks (relevance ≥ 67 percent and mean importance ≥ 3.0) with positive relevance; specialties are averaged into the 6-digit SOC code, and 29 occupations missing ratings receive equal weights. Physical tasks are defined by mapping to the following generalized work activities: 4.A.3.a.1–4 (performing general physical activities, handling and moving objects, controlling machines and processes, operating vehicles and equipment), 4.A.3.b.4–5 (repairing and maintaining equipment) and 4.A.4.a.5 (assisting and caring for others). Live-interaction tasks are defined by 4.A.4.a.4, 4.A.4.a.6, 4.A.4.a.7 and 4.A.4.a.8 (relationships, selling or influencing, negotiating, working with the public), 4.A.4.b.2 to 4.A.4.b.5 (building teams, teaching, directing subordinates, coaching) and 4.A.4.c.2 (staffing). The Dingel–Neiman rule uses the Work Context and Work Activities files of the same release (Section K).

- **OEWS May 2025 (National):** Provides employment and mean annual wage by detailed occupation. Hourly wages are annualized by OEWS at 2,080 hours. Twelve OEWS hybrid codes are mapped explicitly. Occupations lacking O*NET task data (accounting for 1.3 percent of employment) are dropped.
- **Eloundou et al. (2024) labels:** Sourced from github.com/openai/GPTs-are-GPTs and used solely to define the alternative excluded set and for the rank correlation comparison in Section I.
- **External benchmarks:** METR’s cross-domain horizons are taken from github.com/METR/cross-domain-horizon (pinned commit). Published figures from GDPval (win rates by task length and the distribution of task times; [Patwardhan et al., 2025](#)), METR’s studies of maintainer review and of contractor and maintainer times ([Whitfill et al., 2026](#); [Kwa et al., 2025](#)), the Remote Labor Index ([Mazeika et al., 2025](#)), the AEI January 2026 report ([Appel et al., 2026](#)), [Korinek et al. \(2026\)](#) and the fitted curves of [Mertens et al. \(2026\)](#) are recorded, with page references and access dates, in the replication files. They are cited parameters rather than data, and every table that uses one marks it. I read their curves with human time in minutes.

B METR Horizons, Task Suite and Refit

Figure A1 reproduces METR’s headline estimates: the 50 percent time horizon of each model it has measured.

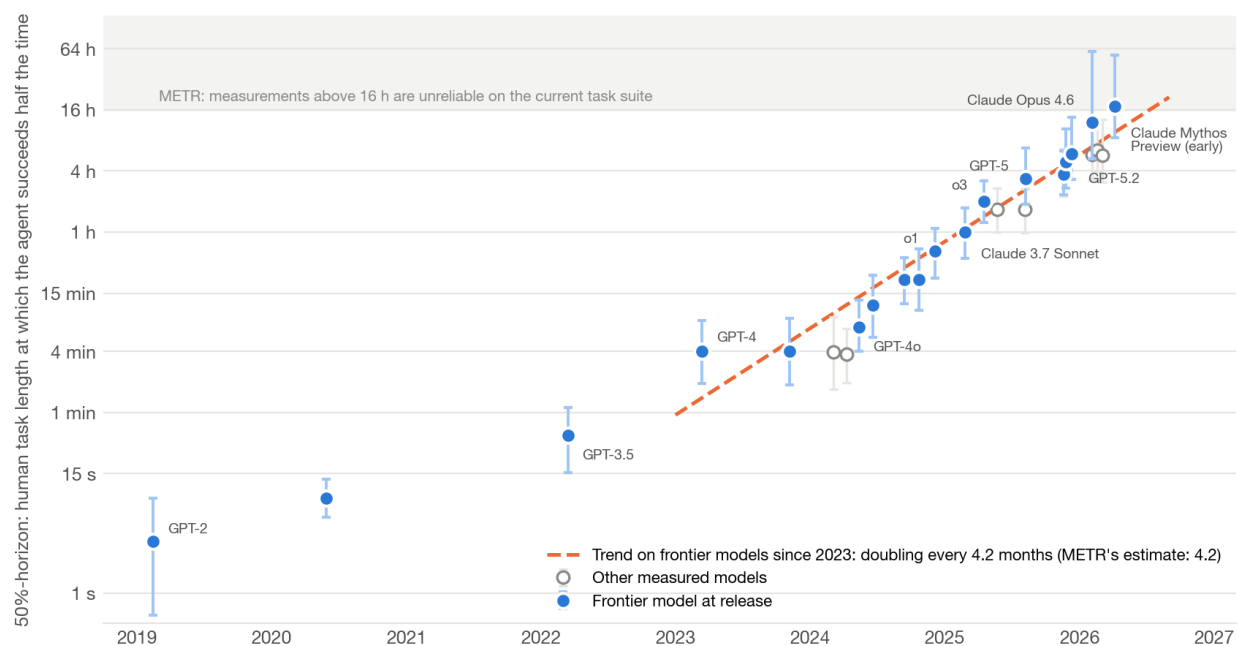


Figure A1: METR’s 50 percent time horizons, reproduced from METR’s published estimates ([METR, 2026](#)). Bars are METR’s published 95 percent confidence intervals. Filled points: frontier model at release. Dashed line: exponential trend fitted to frontier models released from January 2023 with horizons below 16 hours. Grey band: horizons METR describes as unreliable on the current suite.

Table A1 gives the composition of METR’s Time Horizon 1.1 suite by source. Table A2 reports METR’s published 50 and 80 percent horizons alongside the refit, and the fitted slope per doubling of task length, for all models with fitted curves. “Frontier at release” is METR’s `is_sota` indicator; these models are the frontier steps in Figure 2(a). Table A3 reports the fit’s residuals by duration

bin.

Table A1: Composition of METR’s Time Horizon 1.1 task suite. Weight = share of METR’s diversity-adjusted weight in the Mythos Preview (early) fit, where each task in a family of n_f tasks carries weight $1/\sqrt{n_f}$, so a family’s total weight is proportional to $\sqrt{n_f}$. The band from half an hour to eight hours holds 40 percent of the weight on 72 tasks.

Source	Content	Tasks	Families	Human time: median, max	Weight
HCAST	software engineering, machine learning, cybersecurity, and computer-based research tasks	157	69	44 min, 30.0 h	81%
RE-Bench	machine-learning research engineering tasks	5	5	8.0 h, 8.2 h	4%
SWAA	seconds-long software actions (code completion, file selection, request routing)	66	5	5 s, 43 s	15%

Table A2: METR’s published 50 and 80 percent horizons and my refit from the per-task results, all models with fitted curves. Slope = the refit coefficient on $\log_2$ human minutes, sign reversed: the fall in the log-odds of success per doubling of task length. “Frontier at release” is METR’s `is_sota` flag; these models are the frontier steps in Figure 2(a), in date order from GPT-4 (March 2023). Mythos Preview (early) is a pre-release checkpoint measured on 1,149 runs.

Model	Released	p50 METR (min)	p50 refit (min)	p80 METR (min)	p80 refit (min)	Slope	Frontier at release
GPT-4 0314	2023-03-14	4.0	4.0	0.9	0.9	0.64	yes
GPT-4 1106	2023-11-06	4.0	4.0	0.8	0.8	0.59	yes
Claude 3 Opus	2024-03-04	4.0	4.0	0.6	0.6	0.53	
GPT-4 Turbo	2024-04-09	3.7	3.7	0.9	0.9	0.69	
GPT-4o	2024-05-13	7.0	7.0	1.3	1.3	0.56	yes
Claude 3.5 Sonnet (Old)	2024-06-20	11.4	11.4	1.7	1.7	0.50	yes
o1-preview	2024-09-12	20.3	20.3	4.4	4.4	0.63	yes
Claude 3.5 Sonnet (New)	2024-10-22	20.5	20.5	2.6	2.6	0.46	yes
o1	2024-12-05	38.8	38.8	7.1	7.1	0.57	yes
Claude 3.7 Sonnet	2025-02-24	60.4	60.4	12.1	12.1	0.60	yes
o3	2025-04-16	119.7	119.7	30.0	30.0	0.69	yes
Claude 4 Opus	2025-05-22	100.4	100.4	20.4	20.4	0.60	
Claude 4.1 Opus	2025-08-05	100.5	100.5	23.5	23.5	0.66	
GPT-5	2025-08-07	203.0	203.0	38.3	38.3	0.58	yes
Gemini 3 Pro	2025-11-18	224.3	224.3	54.1	54.1	0.68	yes
GPT-5.1-Codex-Max	2025-11-19	223.7	223.7	50.6	50.6	0.65	
Claude Opus 4.5	2025-11-24	293.0	293.0	49.4	49.3	0.54	yes
GPT-5.2	2025-12-11	352.2	352.3	66.0	66.0	0.57	yes
GPT-5.3-Codex	2026-02-05	349.5	349.5	54.7	54.7	0.52	
Claude Opus 4.6	2026-02-05	718.8	718.8	69.9	69.9	0.41	yes
Gemini 3.1 Pro	2026-02-19	384.1	384.1	89.8	89.8	0.66	
GPT-5.4	2026-03-05	341.7	341.7	53.9	53.9	0.52	
Claude Mythos Preview (early)	2026-04-07	1044.8	1044.7	185.9	185.9	0.56	yes

Table A3: Observed minus fitted success by duration bin within METR’s suite, Mythos Preview (early). Observed and fitted are diversity-weighted means over the tasks in the bin; difference in percentage points. Last column: share of the total wage bill in tasks in scope whose AEI duration falls in the bin. Reweighting the differences by the last column gives the change in the expected-success share that would follow if the suite’s residual pattern held in the economy: about -2 points for Mythos Preview (early) and -1 for Claude Opus 4.6.

Human time	METR suite		Mythos Preview (early)			In-scope wage bill in this bin
	Tasks	Weight	Observed success	Fitted	Difference	
< 1 min	67	15%	99%	100%	-0.5	0%
1-15 min	52	24%	99%	98%	+0.3	1%
15-60 min	36	17%	98%	94%	+3.3	6%
1-4 h	18	9%	76%	86%	-9.9	26%
4-16 h	50	31%	68%	67%	+1.3	26%
> 16 h	5	3%	45%	47%	-1.4	0%

C The Excluded Set and Task Examples

Of the 18,796 tasks in the economy, 4,819 are entirely physical and 657 mix physical and other activities; a further 2,579 involve no physical activity but include live interaction. Table A4 illustrates the mapping from task to work activity, and Table A5 gives the wage bill in each class.

Table A4: From task to work activity: the O*NET mapping for one task in each class. Data Entry Keyers: a task in scope and an entirely physical task. General and Operations Managers: a task excluded only for live interaction. Construction Laborers: a task that mixes physical and other activities. A task is excluded if any detailed work activity it maps to sits under a physical or live-interaction generalized work activity.

Task (occupation)	Detailed work activities → generalized work activity	Counted as
Compile, sort, and verify the accuracy of data before it is entered. (Data Entry Keyers)	Compile data or documentation. → Processing Information	in scope
Perform personnel functions, such as selection, training, or evaluation. (General and Operations Managers)	Conduct employee training programs. → Training and Teaching Others (live interaction) Hire personnel. → Staffing Organizational Units (live interaction)	excluded
Measure, mark, or record openings or distances to layout areas where construction work will be performed. (Construction Laborers)	Measure work site dimensions. → Estimating the Quantifiable Characteristics of Products, Events, or Information Mark reference points on construction materials. → Handling and Moving Objects (physical)	excluded
Store completed documents in appropriate locations. (Data Entry Keyers)	Store records or related materials. → Handling and Moving Objects (physical)	excluded

Because the rule records what a task involves rather than where it is done, O*NET files “measure felled logs” under estimating quantities, so log graders and agricultural inspectors pass it. Adding the inspecting work activity to the excluded set moves only 3.2 percent of the wage bill and leaves both occupations in place. Setting is recorded only at the level of the occupation, so as a bound I also apply the rule of [Dingel and Neiman \(2020\)](#), rebuilt on O*NET 30.2, which excludes every task in occupations that cannot be done remotely. Tasks in scope in such occupations carry 25 percent of the wage bill. Dingel and Neiman count school teaching as remote (it could be done at home

Table A5: Task universes. Shares are of the \$10.72 trillion wage bill attached to all tasks.

Universe	Tasks	Share of wage bill
All O*NET tasks in occupations with OEWS employment and wages	18,796	100.0%
In-scope tasks (no physical or live-interaction work activity)	10,741	59.5%
Tasks with an observed AEI duration (June 2026 AEI release, Apr and May 2026)	2,814	25.6%
Computer and Mathematical occupations (in-domain check, SOC 15)	770	5.9%

in 2020), so I also exclude occupations that require close physical contact (Section K). Figure A2 shows the range: the expected-success share under Mythos Preview (early) runs from 24 percent with both exclusions (26 percent with Dingel–Neiman alone) to 52 percent with a lenient rule that excludes a task only if every activity is physical or interactive (Table 1 reports four of these rules over time). The Eloundou et al. (2024) exposure labels give 49 percent. The aggregate results in the main text use the baseline filter; the results for individual occupations in Section 3.4 use both exclusions.

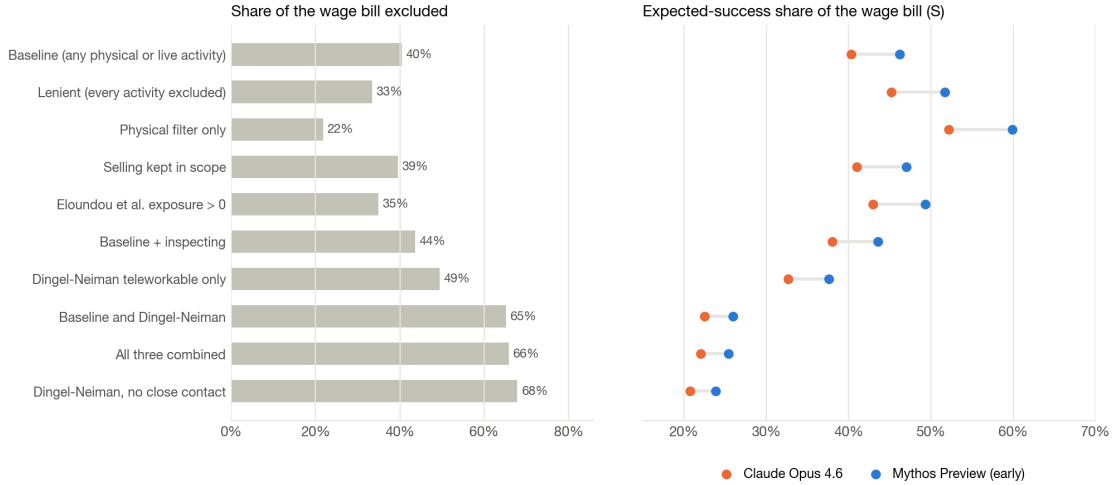


Figure A2: Ten rules for the excluded set. Left: share of the wage bill excluded. Right: expected-success share of the wage bill under both frontier models. Baseline: a task is excluded if any of its work activities is physical or live interaction with people (Section 2.4). Lenient: excluded only if every work activity is. Physical filter only: live interaction kept in scope. Selling kept in scope: generalized work activity 4.A.4.a.6 removed from the interactive set. Eloundou et al.: excluded if the exposure label is zero. Inspecting: tasks with a work activity under generalized work activity 4.A.1.b.2 are also excluded. Dingel–Neiman: every task of an occupation that cannot be done entirely at home is excluded, using their published rule rebuilt on O*NET 30.2 (Section K). No close contact: also excludes occupations whose O*NET physical-proximity rating is 4 or more (at arm’s length or closer), such as school teaching, which Dingel–Neiman count as remote because it can be delivered through a screen.

D Variation within METR’s Suite

Figure A3 plots, for the two frontier models, each METR task’s observed success rate against its human time, with METR’s pooled curve and curves fitted within each task source. Table A6 gives the horizons from each fit and the range of horizons obtained by dropping any one task family and

refitting. The R^2 in the panel titles is the weighted R^2 of task success rates on the pooled curve; the corresponding ceiling if the curve were exactly right, given the number of runs per task, is 0.61 for Claude Opus 4.6 and 0.57 for Mythos Preview (early). Across all 23 models with fitted curves the highest weighted R^2 is 0.68 (GPT-4 0314, ceiling 0.89); the frontier models sit low because most of their tasks are solved nearly every time and the remaining failures are concentrated in particular tasks rather than at particular lengths. The RE-Bench fits are not shown because five tasks do not identify a slope. The SWAA-only horizons are shown but should not be read as measurements: the longest SWAA task takes 43 seconds, so both horizons for Mythos Preview (early) and the 50 percent horizon for Claude Opus 4.6 lie beyond the subsuite’s data, and one percentage point of observed success separates a one-minute from a sixteen-minute fitted horizon.

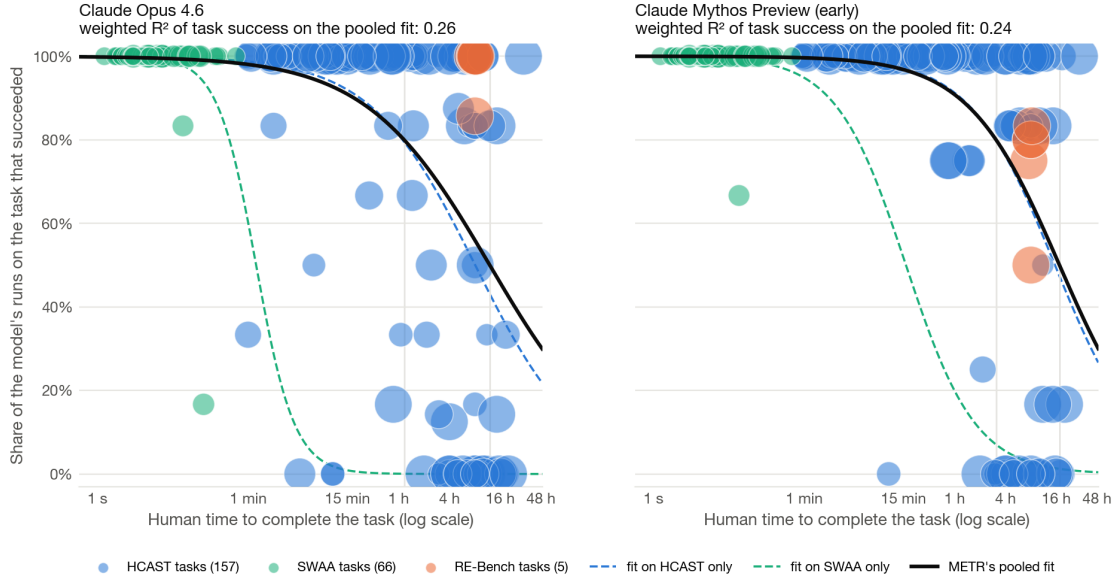


Figure A3: Each point is one METR task; area is proportional to METR’s diversity weight. Black: METR’s pooled fit. Dashed: logistic fits within one task source. Grey lines: the pooled 50 and 80 percent horizons. Left: Claude Opus 4.6. Right: Mythos Preview (early).

Table A6: Horizons from METR’s pooled fit, from fits within one task source, and the range when any one of the 79 task families is dropped. Success = diversity-weighted share of runs that succeeded. †: horizon lies beyond the longest task in the subsuite (an extrapolation, not a reading within the data).

Model	Fit	Tasks	Weight	Success	50% horizon	80% horizon
Claude Opus 4.6	pooled (METR’s fit)	228	100%	79%	12.0 h	1.2 h
	HCAST tasks only	157	81%	74%	8.1 h	1.1 h
	SWAA tasks only	66	15%	98%	1 min [†]	36 s
	RE-Bench tasks only	5	4%	97%	n/a	n/a
	drop any one family				10.3 h to 14.2 h	1.1 h to 1.4 h
Mythos Preview (early)	pooled (METR’s fit)	228	100%	85%	17.4 h	3.1 h
	HCAST tasks only	157	81%	83%	15.4 h	3.1 h
	SWAA tasks only	66	15%	99%	16 min [†]	4 min [†]
	RE-Bench tasks only	5	4%	74%	n/a	n/a
	drop any one family				14.8 h to 20.9 h	2.9 h to 3.6 h

E Task Success and Task Length Outside METR’s Suite

Outside the suite, other success curves can be drawn, but each measures something different from autonomous completion. Two AEI releases report, for each task, the share of conversations that Claude classified as successful alongside their human time. On Claude.ai this share barely changes with duration: across tasks, the fall in the log-odds of success per doubling of human time is between -0.04 and 0.04 , against 0.41 to 0.66 for the METR curves of the Claude models that bracket that traffic. On the API it falls, but more slowly, at 0.12 to 0.21 per doubling within occupations and work activities, and at 0.37 within the same task across the two releases (Section E). The AEI’s authors give reasons to expect these curves to be flatter than METR’s (Appel et al., 2026). Users choose which tasks to bring to Claude, so observed success reflects their judgment of what will work, of the cost of setting up the task and of the time it would save; if they avoid tasks they expect to fail, observed success overstates capability. On Claude.ai, each turn of a conversation lets the user correct course, which breaks a long task into shorter steps. Users may also respond to better models by bringing harder versions of the same task. And the labels are Claude’s own judgment of whether the user’s request was met. GDPval, in which experts grade one-shot deliverables on a fixed set of tasks, avoids the selection of tasks and also shows a flatter gradient, 0.21 per doubling for Claude Opus 4.1 against 0.66 on METR’s suite, but its standard is a comparison with an expert’s work (Patwardhan et al., 2025). Workers’ evaluations of single responses to 6,648 O*NET tasks also give a flatter gradient, 0.08 per doubling pooled over 41 language models and no more than 0.16 for any one of them (Mertens et al., 2026); Section E.1 discusses them. Success also depends on the kind of work, consistent with a “jagged” frontier (Dell’Acqua et al., 2026): METR’s comparison across benchmarks finds that the level of the horizon differs by orders of magnitude between domains, with horizons on visual computer-use tasks 40 to 100 times shorter than on software, although they are growing at similar rates (Kwa and Cheng, 2025). I keep METR’s curves, whose gradient is measured on autonomous completion under controlled conditions, and treat the variation of success across tasks as the least secure part of the measure.

The January and March 2026 AEI releases report, for each O*NET task and platform, the number of conversations that Claude classified as successful or unsuccessful, alongside the mean human time of those conversations. I keep tasks with at least ten classified conversations, which covers about 0.8 million conversations per window and platform, and regress success on $\log_2$ of human minutes in a conversation-weighted binomial model, pooled and with occupation or intermediate work activity fixed effects (Table A7). For tasks observed in both releases, a model with task and release fixed effects asks whether the same task succeeds less often in the release in which its delegated piece is longer: on the API it does, at 0.37 per doubling over 1,411 tasks, while on Claude.ai the slope is -0.08 . Figure A4(a) plots success by duration bin on the API in the February 2026 window against the METR curve of Claude Opus 4.5. For tasks under 15 minutes the curves predict 89 to 96 percent success against 58 to 74 percent observed; for tasks of four to sixteen hours they predict less success than observed on Claude.ai. Across occupations, the rank correlation between the expected-success share of value in scope and observed success is 0.02 to 0.24 .

Using the pooled AEI curves in place of METR’s would not raise the headline share. On the baseline task frame, the API curves give an expected-success share of 25 to 26 percent and the nearly flat Claude.ai curves 40 to 42 percent, against 46 percent under Mythos Preview (early). A flatter gradient raises success on long tasks but lowers it on short ones, and these curves describe the models that served the traffic, not the latest frontier. Against Claude Opus 4.6, the latest of those models (40 percent), the API curves give less and the Claude.ai curves slightly more. Neither is an

alternative estimate, because the labels measure success given delegation rather than autonomous completion.

GDPval reports success by task length on a fixed set of tasks, which users do not select (Patwardhan et al., 2025). Its gold-set tasks were written by professionals, and each deliverable is graded blind against an expert’s. The paper reports win rates for four bins of expert time; I place each bin at its median under a log-normal fitted to the published quartiles and fit a logistic curve to the four rates. The fall in the log-odds per doubling is 0.20 to 0.24 across Claude Opus 4.1, GPT-5 and o3, against 0.58 to 0.69 on METR’s suite for the same models (Figure A4(b)). The tasks are one-shot and well specified, the times are self-reported, and four bins give a coarse gradient. The win rate in this breakdown probably counts only deliverables rated better than the expert’s, whereas the calibration in Section J uses wins and ties, so the gradient and the level rest on slightly different standards.

Figure A4(c) shows the curves that Mertens et al. (2026) fit to workers’ evaluations of single responses, for two Claude models that METR also measured. Their curves are much flatter than METR’s and move up over time rather than out (Section E.1).

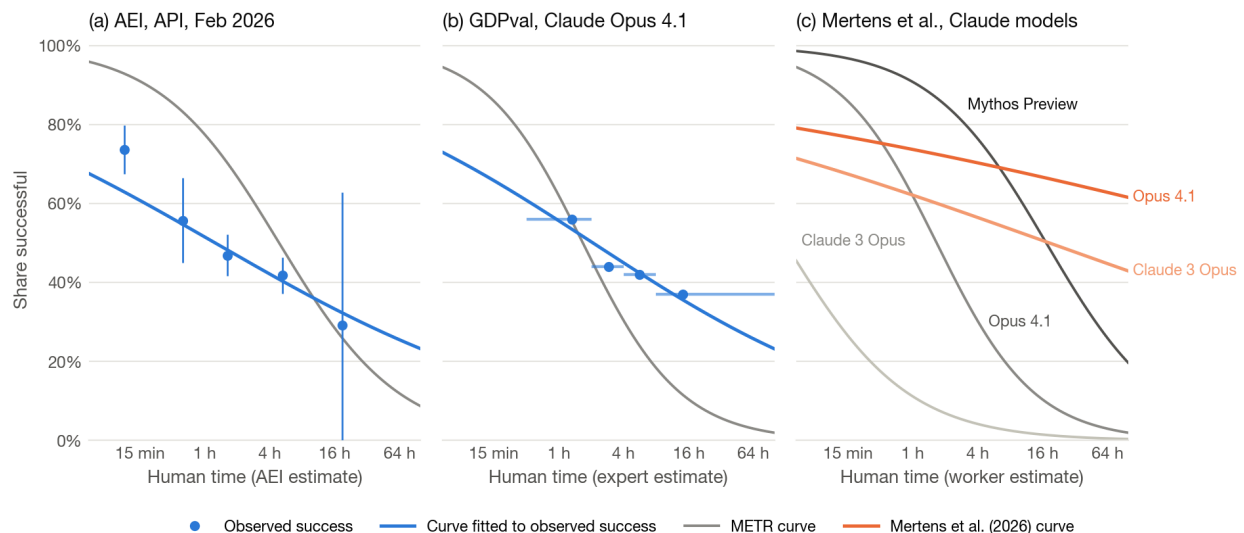


Figure A4: Success against human time outside METR’s suite, against METR’s curves for the same Claude models (grey). Blue: observed success and a logistic curve fitted to it. (a) Share of API conversations in the AEI’s February 2026 window classified as successful, by the task’s mean human time, for tasks with at least ten classified conversations; bars are 95 percent intervals clustered by task. Grey: Claude Opus 4.5, the latest Claude model with a METR curve released before the window (Claude Opus 4.6 was released on its first day). (b) GDPval win rates against an expert for Claude Opus 4.1, by bin of the expert’s time (bars span each bin; points at the bin medians) (Patwardhan et al., 2025). Grey: Claude Opus 4.1. (c) Orange: the curves that Mertens et al. (2026) fit for Claude 3 Opus (March 2024) and Claude Opus 4.1 (August 2025), the share of each model’s responses that workers judge usable without edits, by the worker’s estimate of the task’s length (their Figure A.5, human time in minutes; Section A). Grey: METR’s curves for the same two models and for Mythos Preview (early), released in April 2026, which they did not evaluate (their latest model was released in August 2025). Darker shades are later models. Their task-level points are not public, so only curves are shown. Fall in the log-odds of success per doubling of human time: (a) fitted 0.19, METR 0.54; (b) fitted 0.21, METR 0.66; (c) Mertens et al. (2026) 0.12 and 0.08, METR 0.53, 0.66 and 0.56.

Table A7: AEI task success and human time. Binomial regressions of classified success on $\log_2$ human minutes, weighted by conversations; heteroskedasticity-robust standard errors in parentheses. The slope is reported as the fall in the log-odds of success per doubling, the same units as METR’s slopes (0.41 to 0.66 for the Claude models serving these windows). Within occupation: 6-digit SOC fixed effects. Within activity: fixed effects for the intermediate work activity of the task’s first detailed work activity. Within task: tasks observed in both releases on the same platform (2,225 on Claude.ai, 1,411 on the API), with task and release fixed effects.

Platform	Window	Tasks	Mean success	Fall in log-odds of success per doubling			
				Pooled	Within occupation	Within activity	Within task
Claude.ai	Nov 2025	2,388	68%	0.04 (0.05)	0.10 (0.03)	0.05 (0.04)	
Claude.ai	Feb 2026	2,551	71%	−0.04 (0.07)	0.02 (0.03)	−0.06 (0.04)	−0.08 (0.07)
API	Nov 2025	1,666	49%	0.19 (0.07)	0.12 (0.05)	0.17 (0.05)	
API	Feb 2026	1,680	50%	0.19 (0.09)	0.13 (0.05)	0.21 (0.04)	0.37 (0.06)

E.1 A flatter success curve

Section 5 compares the expected-success share on the curve of [Mertens et al. \(2026\)](#) with the share on METR’s curves. Their models were released between June 2023 and August 2025, ending with GPT-5, so the comparison stops there; I do not extrapolate their curve to Mythos Preview (early). On their curve the threshold shares would say little. At GPT-5’s release, most of the wage bill in scope would sit at success rates between 66 and 73 percent, against 31 to 76 percent on METR’s curve, all of it shorter than their 50 percent horizon and almost none of it shorter than their 80 percent horizon. Their durations are estimated by the same workers who grade the responses.

F Durations, Robustness and Sensitivity

The duration of a task depends on how the work reaches Claude. The June release is the first whose Claude.ai data include Cowork, Anthropic’s agentic desktop product, and for the same task its durations are 1.6 times those in the February release, while API durations barely change. For the same task in the same month, API durations are 0.26 to 0.50 times Claude.ai durations. The earlier releases give an expected-success share of 49 percent (Table A8).

Table A8 changes one assumption at a time, with Mythos Preview (early) as the frontier, in five groups. Across the variants that keep the baseline denominator, other than the in-person rules, the expected-success share ranges from 40 to 60 percent. The duration bridge and the exclusion rule matter most. Doubling every duration gives 40 percent and cuts the 80 percent threshold share to 11 percent; halving them raises that share to 46 percent. Inflating imputed durations alone, which targets the selection of the observed sample rather than the level of the clock, gives 44, 42 and 40 percent at factors of 1.5, 2 and 3. Using the earlier AEI releases for durations raises the share by 3 points. On the curve side, refitting on HCAST tasks only, dropping any one family, switching to a probit link or fixing the asymptote at 0.85 to 0.95 all leave the share within two points of the baseline, although the probit and low-asymptote fits move the 80 percent threshold share to 23 and 20 percent. Task weights within occupations barely matter, including the hours-implied weight. Switching off O*NET’s core filter and weighting every rated task by importance times relevance moves 16 percent of the wage bill onto supplemental tasks but gives 46 percent; restricting that to tasks with importance above 2.5 and relevance above 50 percent gives 46 percent. Dropping wages and weighting by employment gives 44 percent. The scope rules in Table 1 aside, the Eloundou

labels give 49 percent, and crediting mixed tasks' in-scope share at a proportionally shorter duration 48 to 51 percent. The in-person rules of Section 2.4 give 44 percent at the task level and 26 percent with the Dingel–Neiman bound. The two observed-use variants leave the expected-success columns unchanged by construction; what they move is the use column, from 49 to 24 percent under a share threshold and to 44 percent without API traffic. The last two rows restrict the universe to tasks with an observed duration and are not commensurable with the rest; they are the source of the 62 percent that appears if the whole column is read as one range.

Table A8: One assumption changed at a time; frontier = Mythos Preview (early). Denominator: all tasks, except in the last two rows, where it is the observed-duration tasks. Curve variants refit the same model's runs another way (Section A). θ : share of a mixed task's value credited as in scope, evaluated at duration θD_t . The in-person rows are described in Section K.

Variant (one change from baseline)	Expected success	< p80	Observed, cumulative
<i>Duration bridge</i>			
Baseline	46%	27%	49%
Durations from the Nov 2025 and Feb 2026 AEI releases	49%	40%	49%
All durations halved	51%	46%	49%
All durations doubled	40%	11%	49%
Imputed durations only $\times 1.5$	44%	18%	49%
Imputed durations only $\times 2$	42%	15%	49%
Imputed durations only $\times 3$	40%	13%	49%
<i>Success curve (same model, refit)</i>			
Curve fitted on HCAST tasks only	46%	26%	49%
Drop-one-family refit, lowest share (drops <code>mlab</code>)	46%	25%	49%
Drop-one-family refit, highest share (drops <code>spn.cryptanalysis</code>)	47%	31%	49%
Probit link instead of logistic	46%	23%	49%
Logistic with asymptote $\bar{p} = 0.95$ (refit)	46%	30%	49%
Asymptote $\bar{p} = 0.90$	46%	27%	49%
Asymptote $\bar{p} = 0.85$	45%	20%	49%
<i>Task weights and value</i>			
Equal task weights within occupation	46%	27%	49%
Importance $\times$ relevance weights	46%	27%	50%
Importance $\times$ relevance weights, supplemental tasks admitted (no core filter)	46%	26%	47%
Importance $\times$ relevance weights, tasks with importance > 2.5 and relevance > 50%	46%	26%	48%
Frequency-implied weights	45%	28%	51%
Hours-implied weights (duration $\times$ annual frequency)	44%	18%	51%
Employment weights, no wages	44%	29%	45%
<i>Excluded set</i>			
Physical filter only: live-interaction activities in scope (previous baseline)	60%	31%	49%
Selling or influencing others kept in scope	47%	27%	49%
Non-physical set = Eloundou et al. exposure label > 0	49%	24%	49%
Excluded only if every work activity is physical or live interaction (lenient)	52%	31%	49%
Mixed tasks split: share $\theta = 0.25$ of value at duration θD	48%	29%	49%
Mixed tasks split: $\theta = 0.5$	49%	30%	49%
Mixed tasks split: $\theta = 0.75$	51%	30%	49%
Also excluded if any work activity is inspecting (task level)	44%	25%	49%
In-scope set = occupations Dingel and Neiman class as teleworkable	38%	16%	49%
Baseline and Dingel–Neiman teleworkable occupation (occupation-level rule)	26%	12%	49%
and not close contact (O*NET physical proximity < 4)	24%	11%	49%
and not close contact (physical proximity < 4.5)	25%	11%	49%
<i>Observed-use definition (expected success unchanged by construction)</i>			
Observed only if share $\geq 0.01\%$	46%	27%	24%
Observed in Claude.ai only (API traffic left out)	46%	27%	44%
<i>Change of universe (different denominator; not part of the one-change range)</i>			
Observed-duration tasks only (denominator changes)	62%	28%	100%
Observed-duration tasks only, Nov 2025 and Feb 2026 AEI releases	61%	44%	100%

Table A9: Leave-one-out check of the duration imputation. Each observed task is re-imputed from the other observed tasks by the baseline rule, at the level the rule would use for it. $\hat{D}_t/D_t$ is the ratio of imputed to observed duration. Value shares are of the wage bill in observed-duration tasks.

Imputation level	Tasks	Share of observed value	Median $\hat{D}_t/D_t$	Mean log ratio, value-weighted	SD of log ratio	Within factor 2
Detailed work activity	2,736	97%	1.00	+0.02	0.49	87%
Intermediate work activity	69	2%	1.57	+0.94	0.78	57%
Generalized work activity	9	0%	2.61	+1.38	0.84	44%
All observed tasks	2,814	100%	1.00	+0.04	0.50	86%

Table A10: Observed use under alternative definitions. Share of the wage bill in tasks meeting the definition; denominator all tasks. Top block: threshold on the task’s share of Claude.ai or API conversations in a window. Bottom block: number of the seven windows in which the task was published at all on either platform.

Definition of “observed”	Tasks	May 2026 window	Any window, Dec 2024 to May 2026
Published at all (baseline)	6,754	30%	49%
Share of conversations $\geq 0.01\%$	2,667	16%	24%
Share $\geq 0.05\%$	873	6%	10%
Share $\geq 0.1\%$	492	3%	6%
Published in at least 1 of the 7 windows	6,754		49%
Published in at least 2 of the 7 windows	5,783		44%
Published in at least 4 of the 7 windows	3,583		26%
Published in at least 7 of the 7 windows	964		11%

G Decomposing the April 2026 Capability Step

The 80 percent threshold share in Figure 2(a) jumps from 8 to 27 percent when the frontier passes from Claude Opus 4.6 to Mythos Preview (early). The step is a faithful reading of METR’s published fits, but three things make it as large as it is. First, the frontier only moves when a model beats the best 50 percent horizon to date. METR measured two other models between these two: Gemini 3.1 Pro (19 February 2026; 50 per cent horizon 6.4 hours, 80 per cent horizon 1.5 hours) and GPT-5.4 (5 March 2026; 50 per cent horizon 5.7 hours, 80 per cent horizon 0.9 hours). Neither beat the 50 percent horizon of Claude Opus 4.6, so neither appears in the figure, although the 80 percent horizon of Gemini 3.1 Pro was the longer of the two and would have put 10 percent of the wage bill below it; its expected-success share, 38 percent, is below Claude Opus 4.6’s, so the frontier rule and the share agree on the ranking. Second, Claude Opus 4.6 has an unusually flat success curve, 0.41 per doubling against 0.56 for Mythos Preview (early). Its 50 percent horizon is 10 times its 80 percent horizon, against 4 to 6 times for the other frontier models since 2025, so its 80 percent horizon barely moved from that of GPT-5.2 (1.1 to 1.2 hours) while its 50 percent horizon doubled. Much of the 6-point rise in the expected-success share between the two models is therefore the slope steepening in the half-hour-to-eight-hour band rather than the horizon lengthening. Third, the 80 percent threshold share counts tasks shorter than a threshold, and the threshold moved from 1.2 to 3.1 hours, across the part of the duration axis where the wage bill is: 19 percent of the wage bill is in tasks in scope with durations between those two values (Figure 1(b)), and 26 percent is in tasks in scope between one and four hours. The expected-success share integrates the curve instead of counting a threshold, and moves 6 points on the same step.

The threshold is also poorly determined at the top. The 80 percent horizon of Mythos Preview

(early) has a 95 percent interval of 1.6 to 6.6 hours, which puts the share anywhere between 11 and 48 percent. Under Claude Opus 4.6 the interval is 1 to 23 percent. The level of the 80 percent threshold share in 2026 is therefore uncertain by tens of points either way, and the size of the step says as much about where the wage bill sits on the duration axis as about the model.

H Additional Results

Table A11 reports the threshold shares and the other measures behind Section 3 that Table 1 does not. Figure A5(a) splits the May 2026 wage bill by whether a task is shorter than a horizon of Mythos Preview (early) and whether it has appeared in any AEI release. Panel (b) gives the automation share of Claude.ai conversations by the exposure label of Eloundou et al. (2024), for comparison with Figure 2(b).

Table A11: Threshold shares, May 2026, percent. “In scope” = no physical or live-interaction work activity. Durations imputed where unobserved. The fourth row scores excluded tasks by the success curve instead of setting them to zero. SOC 15 (Computer and Mathematical occupations) is an in-domain benchmark.

Share of the wage bill in tasks...	Denominator	Opus 4.6	Mythos Preview
Shorter than the 50 percent horizon	all tasks	58	59
Shorter than the 50 percent horizon	tasks in scope	97	100
Shorter than the 80 percent horizon	tasks in scope	14	45
Expected success, excluded tasks scored by the curve	all tasks	70	79
Shorter than the 80 percent horizon, SOC 15	SOC 15 tasks	3	17
Observed in any AEI release, SOC 15	SOC 15 tasks	82	82

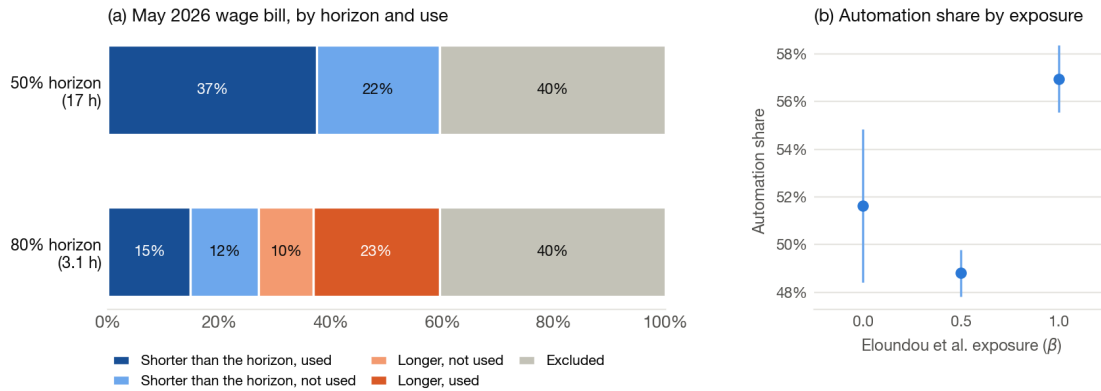


Figure A5: Capability and use: detail. (a) The May 2026 wage bill split by whether a task in scope is shorter than the 50 or 80 percent horizon of Mythos Preview (early) and whether it has a published Claude.ai or API usage share in any AEI release; grey: excluded tasks. (b) Automation share of Claude.ai conversations in May 2026 by the Eloundou et al. (2024) exposure label, for the 2,378 tasks in scope with a published share; bars are 95 percent intervals. Automation: directive and feedback-loop conversations, as a share of conversations with a collaboration pattern (Handa et al., 2025).

H.1 Occupations

Table A12 lists the occupations with the highest expected-success shares under Mythos Preview (early), beside the five largest occupations.

Table A12: The ten occupations with the highest expected-success share of task value under Mythos Preview (early), alongside the five largest occupations by employment. Excluded tasks, and all tasks within occupations that cannot be done remotely (Dingel and Neiman, 2020) or require close physical contact, carry a success probability of zero. Weakest link: the lowest success probability among the occupation’s above-median-importance tasks (evaluating to zero if any such task is excluded). Median duration: wage-weighted baseline durations.

SOC	Occupation	Employment (thousands)	Expected success	Weakest link	Observed use	Median duration (h)
43-4021	Correspondence Clerks	4	85%	74%	85%	1.7
29-2072	Medical Records Specialists	195	82%	0%	71%	1.5
43-9041	Insurance Claims and Policy Processing Clerks	214	81%	69%	67%	3.4
31-9094	Medical Transcriptionists	42	80%	70%	84%	3.9
17-1021	Cartographers and Photogrammetrists	14	79%	71%	44%	2.8
17-3031	Surveying and Mapping Technicians	58	79%	56%	31%	2.5
43-9031	Desktop Publishers	3	79%	65%	78%	3.3
19-1011	Animal Scientists	3	79%	56%	42%	1.2
19-1013	Soil and Plant Scientists	16	78%	56%	22%	3.2
27-3092	Court Reporters and Simultaneous Captioners	13	78%	72%	30%	4.4
31-1120	Home Health and Personal Care Aides	4,306	0%	0%	50%	1.5
41-2031	Retail Salespersons	3,898	0%	0%	73%	2.1
35-3023	Fast Food and Counter Workers	3,854	0%	0%	22%	0.6
11-1021	General and Operations Managers	3,503	23%	0%	44%	7.0
29-1141	Registered Nurses	3,380	14%	0%	45%	2.9

I Comparison with Exposure Measures

Figure A6 plots the correlations discussed in Section 3.4, on the baseline filter. The agreement in panels (a) to (c) is largely the excluded set, which both measures largely respect; panels (d) to (f) condition it away by computing both measures over tasks in scope only. Panel (f) supplies the channel: the occupations the rubric scores as most exposed are the ones whose delegated tasks are longest, and the success curve falls with length. Two cautions: the vertical range in panels (d) and (e) is narrow, so the rank correlation turns on small differences; and the occupations at the top left of panel (e) are mostly occupations with a large excluded share whose small in-scope remainder happens to be short.

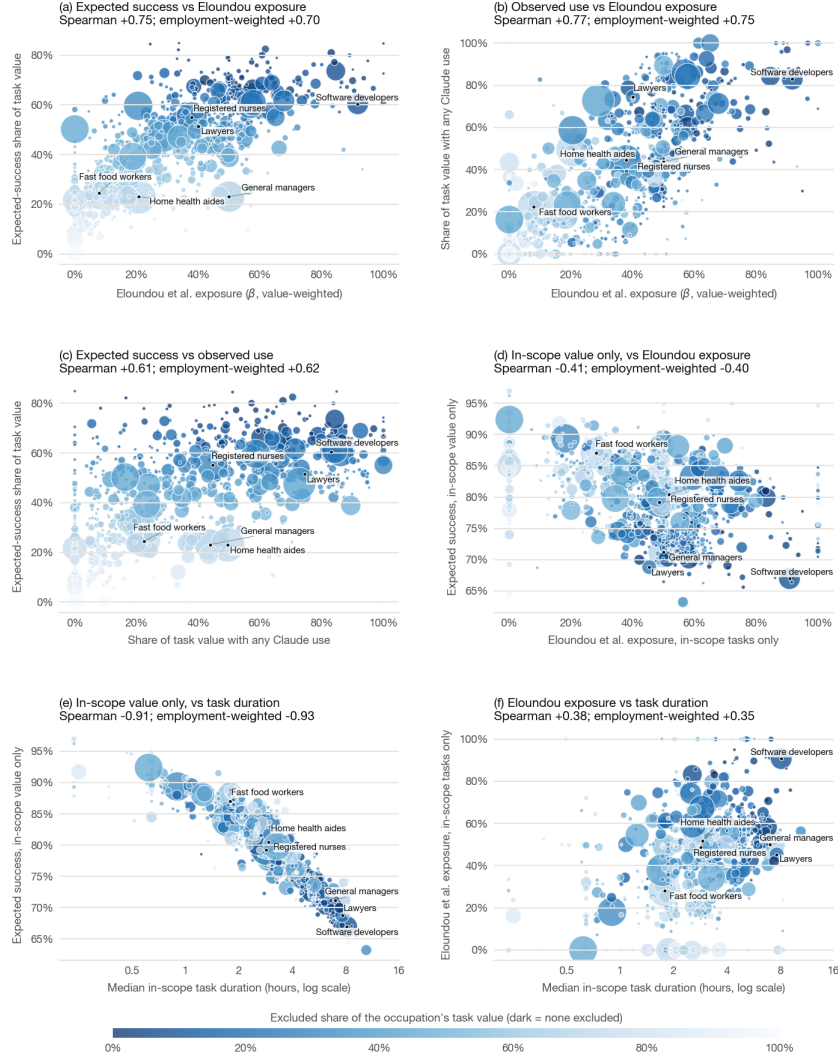


Figure A6: The comparison with rubric-based exposure measures described in Section I, on the baseline filter. One bubble per 6-digit occupation, area proportional to employment, shade the excluded share of the occupation’s task value. Each panel prints the unweighted Spearman rank correlation across the 782 occupations and its employment-weighted Pearson counterpart. Panels (a) to (c) use the additive expected-success share, in which excluded tasks enter at probability zero. Panels (d) to (f) compute both measures over the occupation’s tasks in scope only; the expected-success share is then a function of task duration alone. Panel (f) plots exposure against duration. Eloundou exposure is the value-weighted mean of the task-level β labels within the occupation; median duration is wage-weighted; frontier = Mythos Preview (early).

J External Benchmarks and the Horizon Map

The expected-success share depends on the success curve only through its 50 percent horizon h and slope b , since $P(D) = \sigma(b \log_2(D/h))$ with $b < 0$. Figure A7 maps any (h, b) to the share; every METR model falls on the surface at its own slope. Holding the slope fixed, dividing the horizon by a factor k is equivalent to multiplying every task duration by k , and at METR’s doubling time it corresponds to $\log_2 k$ doubling periods of frontier progress.

Each external benchmark sets its own standard of success. For each, I find the factor k at which

METR’s curve for the same model reproduces the benchmark’s observed success rate at the benchmark’s task lengths, and apply the same factor to the frontier curves (Table A13). GDPval’s gold-set tasks take experts a median of 5 hours, and I treat their durations as log-normal with the published quartiles (Patwardhan et al., 2025). On those lengths METR’s curve for Claude Opus 4.1 predicts 30 percent success, against a win-or-tie rate of 47.6 percent, so the win-or-tie standard implies a longer horizon. The AEI usage horizons come from Appel et al. (2026), who find that Claude succeeds on half of API tasks of about 3.5 hours and half of Claude.ai tasks of about 19 hours, against METR’s 2 hours for the same model. For code judged by repository maintainers rather than by tests, Whitfill et al. (2026) estimate a horizon of about 8 minutes for Claude Sonnet 4.5, against 50 minutes on the automated grader, which they describe as roughly a 7-fold overstatement. The in-context rows apply the 5- to 18-fold gap between METR’s contractors and repository maintainers (Kwa et al., 2025) to every duration. METR’s cross-domain horizons use METR’s own estimates by benchmark (Kwa and Cheng, 2025); several assume a slope of 0.6.

These translations are illustrations, not alternative estimates. Each assumes that the whole gap between two standards acts like a change in task length, with METR’s slope held fixed. If part of the gap is a quality standard that short tasks fail as well, a shift in the level of the curve would fit better and would penalize short tasks less. For the AEI usage horizons, the curves I fit to the AEI’s own success labels do not identify a 50 percent horizon on Claude.ai and place it at 1.0 to 1.3 hours on the API, against 3.5 hours in Appel et al. (2026) (Section E). Those rows combine a published ratio of horizons with METR’s slope, which the labels do not show.

I do not translate the Remote Labor Index (Mazeika et al., 2025). Its five leaderboard entries that match a METR curve complete 1.2 to 4.2 percent of 240 freelance projects, where METR’s curves predict 30 to 51 percent. But the benchmark reports one rate per model and only the mean (28.9 hours) and median (11.5 hours) project length, and 69 percent of its projects are in video, audio, design, CAD and game development, a small part of the wage bill. Read as a shift in the horizon, the gap would also imply that the same agents fail most tasks of an hour, where the AEI labels and GDPval both show success rates of about one half or more.

Table A13: The expected-success share under alternative standards of success and units of human time, as illustrations (see text). Each row divides the frontier model’s 50 percent horizon by the factor shown, holding its slope; months = $\log_2$ of the factor times the doubling time of our fit (4.2 months). †: derived from published figures rather than from data (Section A).

Standard of success or unit of human time	Horizon divided by	Months of progress	Expected success	
			Opus 4.6	Mythos
AEI usage, Claude.ai [†]	0.11	−14	53%	57%
GDPval, wins or ties against experts [†]	0.38	−6	47%	52%
AEI usage, API [†]	0.57	−3	44%	50%
Baseline: METR’s curves as estimated	1	0	40%	46%
All durations doubled	2.0	+4	35%	40%
SWE-bench Verified (METR cross-domain)	3.1	+7	31%	36%
In-context workers, 5× faster (METR) [†]	5.0	+10	27%	31%
Maintainer review, not tests (METR) [†]	7.0	+12	25%	27%
In-context workers, 18× faster (METR) [†]	18	+18	17%	18%
WebArena (METR cross-domain)	21	+18	17%	16%
Computer use, 40× shorter (METR) [†]	40	+23	12%	11%
OSWorld (METR cross-domain)	66	+26	10%	8%
Computer use, 100× shorter (METR) [†]	100	+28	8%	6%

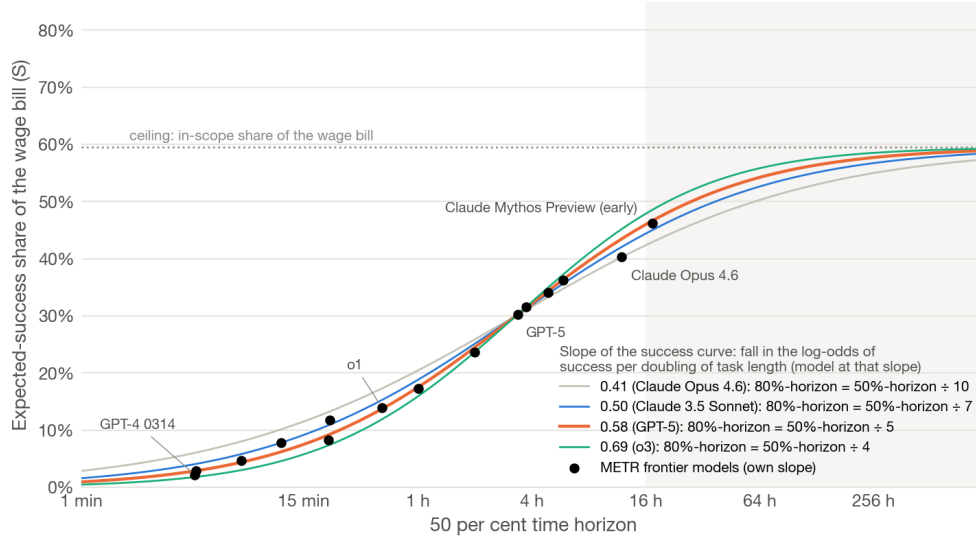


Figure A7: The expected-success share of the wage bill as a function of the 50 percent time horizon, for four values of the slope of METR’s success curve. The slope is the fall in the log-odds of success for each doubling of task length: near the 50 percent horizon, one doubling lowers the probability of success by about a quarter of the slope, 10 to 17 percentage points across these values. It also sets how far below the 50 percent horizon the 80 percent horizon lies: a quarter of it at a slope of 0.69 (o3), a tenth at 0.41 (Claude Opus 4.6). The four values span the frontier models, and each is one frontier model’s own slope. Points: METR’s frontier models at their own horizon and slope. Grey band: horizons METR describes as unreliable. Dotted line: the ceiling set by the in-scope share of the wage bill.

K In-Person Rules

The Dingel–Neiman rule classifies an occupation as teleworkable unless any of fifteen conditions holds, such as rarely using email, working outdoors every day, or rating physical activities, handling objects, dealing with the public or inspecting equipment as very important (Dingel and Neiman, 2020). I transcribe the conditions from their published code and apply them to the Work Context and Work Activities files of O*NET 30.2. On the 720 occupation codes the two O*NET releases share, the rebuilt classification agrees with theirs on 94.6 percent; the rest reflects revised O*NET ratings. On this paper’s weights, teleworkable occupations hold 51 percent of the wage bill and 39 percent of employment, against the 46 and 37 percent that Dingel and Neiman report for 2018. Tasks in occupations without Work Context data (3.7 percent of the wage bill) are kept. For the job-level results I add one condition that their code computes but leaves out of the rule: an O*NET physical-proximity rating (Work Context 4.C.2.a.3) below 4. It excludes school teachers, teaching assistants, childcare workers, food-service supervisors, ushers and gambling dealers, among others; a few exclusions are debatable, such as tutors and air traffic controllers. A threshold of 4.5 gives similar results (Table A8).

L Checking Costs

Suppose that delegating a task to the agent costs the worker checking time vD_t , that a detected failure means the worker redoes the task, and that an undetected or costly failure adds a further cD_t with probability $1 - P_t$. The expected time saved on a delegated task is then $D_t[P_t - v - c(1 - P_t)]$,

and a worker delegates only when this is positive. Table A14 reports the resulting net share of the wage bill saved, for which $v = c = 0$ gives the expected-success share. Anthropic’s reliability adjustment of its own productivity estimate, from 1.8 to 1.0–1.2 percentage points a year (Appel et al., 2026), corresponds to v of 0.23 to 0.35 and a net share of 22 to 31 percent. Only 8.0 percent of the wage bill is in tasks with $P \geq 0.9$ under Mythos Preview (early), and 1.4 percent with $P \geq 0.95$.

Table A14: Net share of the wage bill saved after checking costs. v : checking time as a share of the task’s human time. c : additional cost of a failure as a share of the task’s human time ($c = 0$: failures are caught and redone; $c = 1$: each failure costs one further task). Tasks in scope only; excluded tasks count as zero.

Checking time v (share of the task’s human time)	Mythos Preview (early)		Claude Opus 4.6	
	$c = 0$	$c = 1$	$c = 0$	$c = 1$
0.00	46%	33%	40%	21%
0.10	40%	27%	34%	16%
0.25	31%	18%	25%	9%
0.50	16%	7%	11%	2%

M The Scenarios of Korinek et al.

Table A15 sets this paper’s estimates beside the parameter values of Korinek et al. (2026); Section 4 discusses them. At their parameter values, the static block of their model solved in Figure 4 reproduces their published labor share for 2030 to within 0.1 points in the modest and substantial scenarios and 0.6 points in the extreme scenario, and their net return to capital to within 0.1 points in the modest and substantial scenarios (their Table 3). Output and cognitive employment in that figure are the levels the economy adjusts toward, which run ahead of their published paths for 2030.

Table A15: The parameters of Korinek et al. (2026) beside this paper’s empirical measures. Their values are published figures (their Table 1): the mid-2026 anchor, where provided, and the modest / substantial / extreme values assumed for 2030. Shares are expressed as a fraction of the wage bill; “cognitive occupations” refer to their twelve specified SOC major groups. Expected success is evaluated using METR’s success concept under Mythos Preview (early). The AEI automation share is the usage-weighted share of conversations classified as automation (directive or feedback-loop use) in the June 2026 release.

Dial	Korinek et al. (2026): mid-2026 anchor; 2030 modest / substantial / extreme	This paper
Affected mass m : share of tasks AI can affect	0.14 (Claude usage); 0.2 / 0.3 / 0.5; ceiling 0.624 (all cognitive work)	Expected success in cognitive occupations: 0.36 (0.23 in remote jobs without close contact). Claude use (Claude.ai and API), May 2026: 0.30. Ceiling: 0.60 (all in-scope work). Expected success is a lower bound on m , and closest to ψm with every instance done by AI
Diffusion share d : instances of affected tasks done with AI	0.10 (Census firm survey); 0.2 / 0.4 / 0.6	not measured
Automation share ψ : AI-performed instances automated	0.50 / 0.75 / 0.90, held fixed	AEI, June 2026: 0.49 Claude.ai, 0.94 API. Higher for tasks with a higher success probability, across tasks (Section 3.3)
Gain per AI-performed instance a (log points)	0.30 / 0.45 / 0.80	not measured
Reinstatement ρ : new tasks per automated task	0.5 / 0.25 / 0	not measured